2017-13

Working paper. Economics

ISSN 2340-5031

Quantile Factor Models

Liang Chen, Juan J. Dolado, and Jesús Gonzalo

uc3m | Universidad Carlos III de Madrid

Quantile Factor Models*

Liang Chen[1], Juan J. Dolado[2], and Jesús Gonzalo[3]

[1]*School of Economics, Shanghai University of Finance and Economics, chen.liang@mail.shufe.edu.cn*
[2]*Department of Economics, Universidad Carlos III de Madrid, dolado@eco.uc3m.es*
[3]*Department of Economics, Universidad Carlos III de Madrid, jgonzalo@est-eco.uc3m.es*

April 20, 2019

Abstract

Quantile FactorModels (QFM) represent a new class of factor models for high-dimensional panel data. Unlike Approximate Factor Models (AFM), where only mean-shifting factors can be extracted, QFM also allow to recover unobserved factors shifting other relevant parts of the distributions of observed variables. A quantile regression approach, labeled Quantile Factor Analysis (QFA), is proposed to consistently estimate all the quantile-dependent factors and loadings. Their asymptotic distribution is then derived using a kernel-smoothed version of the QFA estimators. Two consistent model selection criteria, based on information criteria and rank minimization, are developed to determine the number of factors at each quantile. Moreover, in contrast to the conditions required for the use of Principal Components Analysis in AFM, QFA estimation remains valid even when the idiosyncratic errors have heavy-tailed distributions. Three empirical applications (regarding climate, financial and macroeconomic panel data) provide evidence that extra factors shifting quantiles other than the means could be relevant in practice.

Keywords: Factor models, quantile regression, incidental parameters.
JEL codes: C31, C33, C38.

* We are indebted to Ulrich Müller and four anonymous referees for their valuable comments and suggestions. We are also grateful to Dante Amengual, Manuel Arellano, Steve Bond, Guillaume Carlier, Valentina Corradi, Jean-Pierre Florens, Alfred Galichon, Peter R. Hansen, Jerry Hausman, Sophocles Mavroeidis, Bent Nielsen, Enrique Sentana, Liangjun Su, and participants at several seminars and workshops for many helpful discussions. Financial support from the National Natural Science Foundation of China (Grant No.71703089), The Open Society Foundation, The Oxford Martin School, the Spanish Ministerio de Economía y Competitividad (grants ECO2016-78652 and Maria de Maeztu MDM 2014-0431), Bank of Spain (ER grant program), and MadEco-CM (grant S205/HUM-3444) is gratefully acknowledged. The usual disclaimer applies

1 Introduction

Following the key contributions by Ross (1976), Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986) to the theory of approximate factor models (AFM henceforth) in the context of asset pricing, the analysis and applications of this class of models have proliferated thereafter. As is well known, AFM imply that a panel X_{it} of N variables (units), each with T observations, has a factor-structure representation given by: $X_{it} = \lambda_i' f_t + \epsilon_{it}$, where $\lambda_i = [\lambda_{1i}, .., \lambda_{ri}]'$ and $f_t = [f_{1t}, .., f_{rt}]'$ are $r \times 1$ vectors of factor loadings and common factors, respectively, with $r \ll N$, and ϵ_{it} are zero-mean weakly dependent idiosyncratic disturbances which are uncorrelated with the factors.

The fact that it is easy to construct theories involving common factors, at least in a narrative version, together with the availability of simple estimation procedures for AFM — being Principal Components Analysis (PCA hereafter) the most popular, has led to their extensive use in many fields of economics.[1] More recently, a conventional characterization of cross-sectional dependence among error terms in Panel Data has relied on the use of a finite number of unobserved common factors. These originate from economy-wide shocks that affect all units with different intensities (loadings), in addition to idiosyncratic (individual-specific) disturbances. Interactive fixed-effects models can be easily estimated by PCA (see Bai 2009) or by common correlated effects (see Pesaran 2006), and there are even generalizations of these techniques dealing with nonlinear panel single-index models (see Chen et al. 2018). Likewise, the surge of Big Data technologies has made factor models a key tool in dimension reduction and predictive analytics for very large datasets (see Diebold 2012 for a survey).

Our departure point in this paper is to notice that the standard regression interpretation of a static AFM as a linear conditional mean model of X_{it} given f_t, that is, $\mathbb{E}(X_{it}|f_t) = \lambda_i' f_t$, entails two possibly restrictive features. First, PCA does not capture hidden factors that may shift characteristics (moments or quantiles) of the distribution of X_{it} other than the means. Second, neither the loadings λ_i nor the factors f_t are allowed to vary across the distributional characteristics of each unit in the panel.

A simple way of illustrating the limitations of the conventional formulation of AFM is to consider the factor structure in the following simple *location-scale shift* model: $X_{it} = \alpha_i f_{1t} + f_{2t}\epsilon_{it}$, with $f_{1t} \neq f_{2t}$ (both are scalars), $f_{2t} > 0$ and $\mathbb{E}(\epsilon_{it}) = 0$, such that the first factor (f_{1t}) shifts location, whereas the second one (f_{2t}) shifts scale[2]. This model can be rewritten in quantile-regression (QR, hereafter) format as $X_{it} = \lambda_i'(\tau) f_t + u_{it}(\tau)$, with $0 < \tau < 1$, $\lambda_i(\tau) =$

[1] See, *inter alia*, Bai and Ng (2008b) and Stock and Watson (2011). Early applications of AFM abound in Aggregation Theory, Consumer Theory, Business Cycle Analysis, Finance, Monetary Economics, and Monitoring and Forecasting, among others.

[2] This model is further discussed in subsection 2.2 below, alongside other illustrative models representing potential factor structures of X_{it}.

$[\alpha_i, \mathsf{Q}_\epsilon(\tau)]'$, $f_t = [f_{1t}, f_{2t}]'$, $u_{it}(\tau) = f_{2t}[\epsilon_{it} - \mathsf{Q}_\epsilon(\tau)]$, where $\mathsf{Q}_\epsilon(\tau)$ represents the quantile function of ϵ_{it}, and the conditional quantile $\mathsf{Q}_{u_{it}(\tau)}[\tau|f_t] = 0$.[3] PCA will only extract the location-shifting factor f_{1t} in this model, but it will fail to capture the scale-shifting factor f_{2t} and the quantile-dependent loadings $\lambda_i(\tau)$ in its QR representation. Also notice that, when the distribution of ϵ_{it} is symmetric, then f_t could be considered as being quantile dependent, i.e., $f_t(\tau)$, since $f_t(\tau) = f_{1t}$ for $\tau = 0.5$, and $f_t(\tau) = [f_{1t}, f_{2t}]'$ for $\tau \neq 0.5$. Together with other examples discussed in subsection 2.2 further below, this means that the general class of models to be considered in the sequel would be one where the loadings, the factors and the number of factors are all simultaneously allowed to be quantile-dependent objects, namely, $\lambda_i(\tau), f_t(\tau) \in \mathbb{R}^{r(\tau)}$ for $\tau \in (0, 1)$. In what follows, we denote this class of models as *Quantile Factor Models* (QFM, hereafter), whose detailed definition is provided in Section 2 below.

That said, our goal in this paper is to develop a common factor methodology for QFM which is flexible enough to capture those quantile-dependent objects which standard AFM tools may be unable to recover. To do so, we analyze their estimation and inference, including the selection of the number of factors at each quantile τ. In a nutshell, QFM could be thought of as capturing the same type of flexible generalization that QR techniques represent for linear regression models.

To help understand how this new methodology works, we first propose an estimation approach for the quantile-dependent objects in QFM, labeled *Quantile Factor Analysis* (QFA, henceforth). The QFA estimation procedure relies on the minimization of the standard *check* function in QR (instead of the standard quadratic loss function used in AFM) to jointly estimate the common factors $f_t(\tau)$ and the loadings $\lambda_i(\tau)$ at a given quantile τ. However, since the objective function for QFM is not convex in the relevant parameters, we introduce an iterative QR algorithm that yields estimators of the quantile-dependent objects. We then derive their average rates of convergence, and propose two consistent selection criteria, based on information criteria and rank minimization, to choose the number of factors at each τ. In addition, we establish asymptotic normality for QFA estimators based on smoothed QR (see e.g., Horowitz 1998 and Galvao and Kato 2016). Lastly, our asymptotic results and the proposed selection criteria provide guidance on how to discriminate between AFM and QFM structures.

The key contributions of our paper to the literature on FM can be summarized as follows:

1. We propose a new class of factor structures, QFM, provide an estimation method, QFA, of the underlying quantile-dependents objects in QFM, and characterize the asymptotic properties of such estimators. In particular, we show that the average convergence rates of the QFA estimators are the same as the PCA estimators of Bai and Ng (2002), which is a crucial result for showing the consistency of the two selection criteria used to estimate the number of factors at each τ. In addition, similar to Bai (2003), our QFA estimators

[3] Throughout the paper we use $Q_W[\tau|Z]$ to denote the conditional quantile of W given Z.

based on smoothed QR are shown to converge at the parametric rates ($\sqrt{N}$ and $\sqrt{T}$) to normal distributions.

2. The problems of incidental parameters and non-smooth object functions require innovative strategies to derive all the above-mentioned results. This leads to the use in our proofs of some novel techniques borrowed from the theory of empirical processes. Moreover, our proof strategy can be easily extended to some other nonlinear factor models (e.g., probit and logit factor models considered by Chen et al. 2018) with smooth object functions.
3. The QFA estimators inherit from QR certain robustness properties to the presence of outliers and heavy-tailed distributions in the idiosyncratic component of a factor model which render PCA invalid. In effect, while PCA requires the idiosyncratic errors to have eighth bounded moments, QFA only needs the existence and smoothness of the density function. Thus, at $\tau = 0.5$, QFA can be viewed as a robust alternative to PCA.
4. The extra factors obtained by the QFA estimation procedure can be used to improve the monitoring and forecasting performance of variables in a factor-augmented regression setup, as well as to facilitate the factor identification process, depending on the application at hand. For instance, in finance these "new" factors could be interpreted as volatility or tail-risk factors driving assets returns; with income data, they could represent common factors behind income inequality; and with climate data they could represent common features behind global extreme temperatures at both tails of their distribution, etc.

Related literature

There is a recent literature that attempts to make the AFM setup more flexible. For example, Su and Wang (2017) allow the factor loadings to be time varying, while Pelger and Xiong (2018) allow them to be state dependent. Chen et al. (2009) provide a theory for nonlinear principal components, where they suggest using sieve estimation to retrieve nonlinear factors. Finally, Gorodnichenko and Ng (2017) propose an algorithm to estimate level and volatility factors simultaneously. Different from these studies, our approach to modelling nonlinearities in factor models is through the conditional quantiles of the observed variables.

There is also a growing literature on heterogeneous panel quantile models with factor structures, especially in financial economics. The main idea is that a few unobservable factors explain co-movements of asset return distributions in a large range of asset returns observed at high frequencies, as in stock markets. In parallel and independent research, we have recently come across two related studies to ours which focus on similar issues.[4] First, Ma et al. (2017) propose estimation and inference procedures in semiparametric quantile factor models, in which factor loadings/betas are smooth functions of a small number of observables under the assumption that the included factors all have non zero mean. Then, sieve techniques are used to obtain

[4]We only became aware of these two papers after the working paper version of our study, referred to in the sequel as Chen et al. (2017), had been submitted

preliminary estimation of these functions for each time period; next the factor structure is imposed in a sequential fashion to estimate the factor returns by GLS under weak conditions on cross-sectional and temporal dependence. We depart from these authors in that we do not need to assume the loadings to depend on observables and, foremost, in that not only loadings but also factors are quantile-dependent objects in our setup.

Second, in a closely related paper, Ando and Bai (2018) (AB 2018, hereafter) use a setup similar to ours where the unobservable factor structure is also allowed to be quantile dependent. They use Bayesian MCMC and frequentist estimation approaches, the latter building upon our iterative procedure, as it is duly acknowledged in their paper. However, we differ from AB (2018) in several respects which could make our QFA approach valuable: (i) our assumptions are less restrictive, since we rely on properties of the density, as in QR, while AB (2018) needs all the moments of the idiosyncratic errors to exist, (ii) the proofs of the main results are also noticeably different since we believe that our proof strategy can solve some potential problems appearing in theirs, (iii) our rank-minimization selection criterion to estimate the number of factors is computationally more efficient and exhibits a better finite-sample performance than the information-criteria-based method considered by AB (2018).

Lastly, it is also worth highlighting that the illustrative location-scale shift model above, where $f_{1t} \neq f_{2t}$, is behind a current line of research in asset pricing which has been coined the "idiosyncratic volatility puzzle" by Ang et al. (2006). This approach focuses on the co-movements in the idiosyncratic *volatilities* of a panel of asset returns, and basically consists of applying PCA to the squared residuals, once the PCA mean-shifting factors have been removed from the data (a procedure labeled PCA-SQ, hereafter).[5] For example, this technique would fit perfectly to the illustrative example discussed above. Yet, while the QFA estimation approach is able to recover the whole factor structure for more general models than the previous one (see subsection 2.2) or when the idiosyncratic errors lack bounded eighth moments, PCA-SQ fails to do so. Hence, to the best of our knowledge, QFA becomes the first estimation procedure capable of dealing with these issues.

Structure of the Paper. The rest of the paper is organized as follows. Section 2 defines QFM and provides a list of simple illustrative examples where the new QFM methodology applies. In Section 3, we present the QFA estimator and its computational algorithm, establish the average rates of convergence of all the quantile-dependent objects, and propose two consistent selection criteria to choose the number of factors at each quantile, which help when discriminating between AFM and QFM. Section 4 introduces a kernel-smoothed version of the QFA estimators to derive their asymptotic distributions. Section 5 contains some Monte Carlo simulation results to evaluate the performance in finite samples of our estimation procedures relative to other

[5] See, e.g., Barigozzi and Hallin 2016, Herskovic et al. 2016 and Renault et al. 2017. Notice that the volatility co-movement does not arise from omitted factors in the AFM but from assuming a genuine factor structure in the idiosyncratic volatility processes.

alternative approaches under different assumptions about the idiosyncratic error terms. Section 6 considers several empirical applications using three large panel datasets, where we document the relevance of factors shifting other moments of the distributions of the data rather than just their means. Finally, Section 7 concludes and suggests several avenues for further research. Proofs of the main results are collected in the online appendix.

Notations: We use $\|\cdot\|$ to denote the Frobenius norm. For a matrix A with real eigenvalues, let $\rho_j(A)$ denote the jth largest eigenvalue. Following Van der Vaart and Wellner (1996), the symbol $\lesssim$ means "left side bounded by a positive constant times the right side" (the symbol $\gtrsim$ is defined similarly), and $D(\cdot, g, \mathcal{G})$ denotes the packing number of space $\mathcal{G}$ endowed with metric g.

2 The Model and Some Examples

This section starts by introducing the main definitions to be used throughout the paper. Next, we show how to derive the QFM representation of several illustrative DGPs exhibiting different factor structures.

2.1 Quantile Factor Models

Suppose that the observed variable X_{it}, with $i = 1, .., N$ and $t = 1, ..., T$, has the following QFM structure:

$$X_{it} = \lambda_i'(\tau) f_t(\tau) + u_{it}(\tau), \text{ for } \tau \in (0, 1), \tag{1}$$

where the common factors $f_t(\tau)$ is a $r(\tau) \times 1$ vector of unobservable random variables, $\lambda_i(\tau)$ is a $r(\tau) \times 1$ vector of fixed factor loadings, and the idiosyncratic error $u_{it}(\tau)$ is assumed to satisfy the following quantile restriction:

$$P[u_{it}(\tau) \leq 0 | f_t(\tau)] = \tau.$$

Alternatively, (1) implies that

$$\mathsf{Q}_{X_{it}}[\tau | f_t(\tau)] = \lambda_i'(\tau) f_t(\tau),$$

where the factors, the loadings, and the number of factors are all allowed to be quantile-dependent.

2.2 Examples

In this section we provide a few illustrative examples of QFMs derived from different specifications of location-scale shift models and related ones. By means of these simple illustrations, the objective is to show that there are instances where the standard AFM methodology fails to capture the full factor structure and therefore requires the use of our alternative QFM approach.

Example 1. ***Location-shift model.*** *$X_{it} = \alpha_i f_{1t} + \epsilon_{it}$, where $\{\epsilon_{it}\}$ are zero-mean i.i.d errors independent of f_{1t} with cumulative distribution function (CDF) F_ϵ. Let $\mathsf{Q}_\epsilon(\tau) = \mathsf{F}_\epsilon^{-1}(\tau) = \inf\{c : \mathsf{F}_\epsilon(c) \leq \tau\}$ be the quantile function of ϵ_{it}. Moreover, assume that the median of ϵ_{it} is 0, i.e., $\mathsf{Q}_\epsilon(0.5) = 0$, then this simple model has a QFM representation (1) by defining $\lambda_i(\tau) = [\mathsf{Q}_\epsilon(\tau), \alpha_i]'$, $f_t(\tau) = [1, f_{1t}]'$ for $\tau \neq 0.5$, and $\lambda_i(\tau) = \alpha_i$, $f_t(\tau) = f_{1t}$ for $\tau = 0.5$. However, note that the standard estimation method (PCA) for this AFM may not be consistent if the distribution of ϵ_{it} has heavy tails. For example, Assumption C of Bai and Ng (2002) requires $\mathbb{E}[\epsilon_{it}^8] < \infty$, which is not satisfied if, e.g. ϵ_{it} follows the standard Cauchy or some Pareto distributions .*

Example 2. ***Location-scale shift model (same sign-restricted factor).*** *$X_{it} = \alpha_i f_{1t} + f_{1t}\epsilon_{it}$, where $f_{1t} > 0$ for all t and $\{\epsilon_{it}\}$ are defined as in Example 1. This model has a QFM representation (1) by defining $\lambda_i(\tau) = \mathsf{Q}_\epsilon(\tau) + \alpha_i$ and $f_t(\tau) = f_{1t}$ for all τ, such that the loadings of the factor f_{1t} are the only quantile-dependent objects.*

Example 3. ***Location-scale shift model (different factors).*** *$X_{it} = \alpha_i' f_{1t} + (\eta_i' f_{2t})\epsilon_{it}$, where $\{\epsilon_{it}\}$ are defined as in Example 1, $\alpha_i, f_{1t} \in \mathbb{R}^{r_1}$, η_i, $f_{2t} \in \mathbb{R}^{r_2}$, and $\eta_i' f_{2t} > 0$, such that f_{jt} $(j = 1, 2)$ are vectors of r_j factors. When f_{1t} and f_{2t} do not share common elements, this model has a QFM representation (1) with $\lambda_i(\tau) = [\alpha_i', \eta_i' \mathsf{Q}_\epsilon(\tau)]'$, $f_t(\tau) = [f_{1t}', f_{2t}']'$ for $\tau \neq 0.5$, and $\lambda_i(\tau) = \alpha_i$, $f_t(\tau) = f_{1t}$ for $\tau = 0.5$.*

Example 4. ***Location-scale shift model with two idiosyncratic errors.*** *$X_{it} = \alpha_i f_{1t} + f_{2t}\epsilon_{it} + f_{3t}e_{it}$, where ϵ_{it} and e_{it} are two independent normal random variables with variances σ_ϵ^2 and σ_e^2. This model is observationally equivalent to $X_{it} = \alpha_i f_{1t} + \sqrt{f_{2t}^2\sigma_\epsilon^2 + f_{3t}^2\sigma_e^2} \cdot v_{it}$ where v_{it} follows a standard normal distribution. Thus, it has a QFM representation (1) with $\lambda_i(\tau) = [\alpha_i, \Phi^{-1}(\tau)]'$, $f_t(\tau) = [f_{1t}, \sqrt{f_{2t}^2\sigma_\epsilon^2 + f_{3t}^2\sigma_e^2}]'$ for $\tau \neq 0.5$, and $\lambda_i(\tau) = \alpha_i$, $f_t(\tau) = f_{1t}$ for $\tau = 0.5$, where Φ^{-1} is the quantile function of the standard normal distribution.*

Example 5. ***Location-scale shift model with an idiosyncratic error and its cube.*** *$X_{it} = \alpha_i f_{1t} + f_{2t}\epsilon_{it} + c_i f_{3t}\epsilon_{it}^3$, where ϵ_{it} is a standard normal random variable. Let f_{2t}, f_{3t}, c_i be positive, then X_{it} has an equivalent representation in form of (1) with $\lambda_i(\tau) = [\alpha_i, \Phi^{-1}(\tau), c_i\Phi^{-1}(\tau)^3]'$, $f_t(\tau) = [f_{1t}, f_{2t}, f_{3t}]'$ for $\tau \neq 0.5$, and $\lambda_i(\tau) = \alpha_i$, $f_t(\tau) = f_{1t}$ for $\tau = 0.5$. In particular, if $c_i = 1$ for all i and noticing that the mapping $\tau \mapsto \Phi^{-1}(\tau)^3$ is strictly increasing, then we have for $\tau \neq 0.5$, $Q_{X_{it}}[\tau|f_t(\tau)] = \alpha_i f_{1t} + \Phi^{-1}(\tau) \cdot [f_{2t} + f_{3t}\Phi^{-1}(\tau)^2]$, so that there exists a QFM representation (1) with $\lambda_i(\tau) = [\alpha_i, \Phi^{-1}(\tau)]'$ and $f_t(\tau) = [f_{1t}, f_{2t} + f_{3t}\Phi^{-1}(\tau)^2]'$ for $\tau \neq 0.5$.*

Notice that in this case, the second factor in $f_t(\tau)$, $f_{2t} + f_{3t}\Phi^{-1}(\tau)^2$, is quantile dependent even for $\tau \neq 0.5$.

Not surprisingly, PCA only works in Example 1, which corresponds to an AFM, when the idiosyncratic errors satisfy certain moment conditions. In all the remaining cases, PCA will only yield consistent estimates of the location-shift factors; however, it will fail to capture those extra factors which shift quantiles other than the means (Examples 3, 4 and 5) and, even if it extracts all relevant factors, it will miss their corresponding quantile-varying loadings (Example 2). In the sequel, we will therefore propose QFA as a new estimation procedure to estimate both sets of quantile-dependent objects in QFM.

3 Estimators and Their Asymptotic Properties

To simplify the notations, we suppress hereafter the dependence of $f_t(\tau), \lambda_i(\tau), r(\tau)$ and $u_{it}(\tau)$ on τ, so that the QFM in (1) is rewritten as:

$$X_{it} = \lambda_i' f_t + u_{it}, \quad P[u_{it} \leq 0 | f_t] = \tau, \tag{2}$$

where $\lambda_i, f_t \in \mathbb{R}^r$. Suppose that we have a sample of observations $\{X_{it}\}$ generated by (2) for $i = 1, \ldots, N$, and $t = 1, \ldots, T$, where the realized values of $\{f_t\}$ are $\{f_{0t}\}$ and the true values of $\{\lambda_i\}$ are $\{\lambda_{0i}\}$. We take a fixed-effects approach by treating $\{\lambda_{0i}\}$ and $\{f_{0t}\}$ as parameters to be estimated. In Section 3.1, we consider the estimation of $\{\lambda_{0i}\}$ and $\{f_{0t}\}$ while r is assumed to be known. Section 3.2 deals with the estimation of r for each quantile. Finally, in Section 3.3 we discuss how to discriminate between AFM and QFM based the estimated number of mean and quantile factors.

3.1 Estimating Factors and Loadings

It is well known in the literature on factor models that $\{\lambda_{0i}\}$ and $\{f_{0t}\}$ cannot be separately identified without imposing normalizations (see Bai and Ng 2002). Without loss of generality, we choose the following normalizations:

$$\frac{1}{T}\sum_{t=1}^{T} f_t f_t' = \mathbb{I}_r, \quad \frac{1}{N}\sum_{i=1}^{N} \lambda_i \lambda_i' \text{ is diagonal with non-increasing diagonal elements.} \tag{3}$$

Let $M = (N+T)r$, $\theta = (\lambda_1', \ldots, \lambda_N', f_1', \ldots, f_T')'$, and $\theta_0 = (\lambda_{01}', \ldots, \lambda_{0N}', f_{01}', \ldots, f_{0T}')'$ denotes the vector of true parameters, where we also suppress the dependence of θ and θ_0 on M

to save notation. Let $\mathcal{A}, \mathcal{F} \subset \mathbb{R}^r$ and define:

$$\Theta^M = \left\{\theta \in \mathbb{R}^M : \lambda_i \in \mathcal{A}, f_t \in \mathcal{F} \text{ for all } i, t, \{\lambda_i\} \text{ and } \{f_t\} \text{ satisfy the normalizations in (3)}\right\}.$$

Further, define:

$$\mathbb{M}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau(X_{it} - \lambda_i' f_t)$$

where $\rho_\tau(u) = (\tau - \mathbf{1}\{u \le 0\})u$ is the check function. The QFA estimator of θ_0 is defined as:

$$\hat{\theta} = (\hat{\lambda}_1', \dots, \hat{\lambda}_N', \hat{f}_1', \dots, \hat{f}_T')' = \arg\min_{\theta \in \Theta^M} \mathbb{M}_{NT}(\theta).$$

It is obvious that the way in which our estimator is related to the PCA estimator studied by Bai and Ng (2002) and Bai (2003) is analogous to how standard least-squares regressions are related to QR. However, unlike Bai (2003)'s PCA estimator, our estimator $\hat{\theta}$ does not yield an analytical closed form. This makes it difficult not only to find a computational algorithm that would yield the estimator, but also the analysis of its asymptotic properties. In the sequel, we introduce a computational algorithm called *iterative quantile regression* (IQR, hereafter) that can effectively find the stationary points of the object function. In parallel, Theorem 1 shows that $\hat{\theta}$ achieves the same convergence rate as the PCA estimators for AFM.

To describe the algorithm, let $\Lambda = (\lambda_1, \dots, \lambda_N)'$, $F = (f_1, \dots, f_T)'$, and define the following averages:

$$\mathbb{M}_{i,T}(\lambda, F) = \frac{1}{T}\sum_{t=1}^{T}\rho_\tau(X_{it} - \lambda' f_t) \quad \text{and} \quad \mathbb{M}_{t,N}(\Lambda, f) = \frac{1}{N}\sum_{i=1}^{N}\rho_\tau(X_{it} - \lambda_i' f).$$

Note that we have $\mathbb{M}_{NT}(\theta) = N^{-1}\sum_{i=1}^{N}\mathbb{M}_{i,T}(\lambda_i, F) = T^{-1}\sum_{t=1}^{T}\mathbb{M}_{t,N}(\Lambda, f_t)$. The main difficulty in finding the global minimum of $\mathbb{M}_{NT}$ is that this object function is not convex in θ. However, for given F, $\mathbb{M}_{i,T}(\lambda, F)$ happens to be convex in λ for each i and likewise, for given Λ, $\mathbb{M}_{t,N}(\Lambda, f)$ is convex in f for each t. Thus, both optimization problems can be efficiently solved by various linear programming methods (see Chapter 6 of Koenker 2005). Based on this observation, we propose the following iterative procedure:

Iterative quantile regression (IQR):

Step 1: Choose random starting parameters: $F^{(0)}$.

Step 2: Given $F^{(l-1)}$, choose $\lambda_i^{(l-1)} = \arg\min_\lambda \mathbb{M}_{i,T}(\lambda, F^{(l-1)})$ for $i = 1, \dots, N$; given $\Lambda^{(l-1)}$, choose $f_t^{(l)} = \arg\min_f \mathbb{M}_{t,N}(\Lambda^{(l-1)}, f)$ for $t = 1, \dots, T$.

Step 3: For $l = 1, \dots, L$, iterate the second step until $\mathbb{M}_{NT}(\theta^{(L)})$ is close to $\mathbb{M}_{NT}(\theta^{(L-1)})$, where $\theta^{(l)} = (\text{vech}(\Lambda^{(l)})', \text{vech}(F^{(l)})')'$.

Step 4: Normalize $\Lambda^{(L)}$ and $F^{(L)}$ so that they satisfy the normalizations in (3).

To see the connection between the IQR algorithm and the PCA estimator proposed by Bai (2003), suppose that $r = 1$, and replace the check function in the IQR algorithm by the least-squares loss function. Then, it is easy to show that the second step of the algorithm above yields $\Lambda^{(l-1)} = (X'F^{(l-1)})/\|F^{(l-1)}\|^2$ and $F^{(l)} = (X\Lambda^{(l-1)})/\|\Lambda^{(l-1)}\|^2 = XX'F^{(l-1)}/C_{l-1}$, where X is the $T \times N$ matrix with elements $\{X_{it}\}$, and $C_l = \|F^{(l)}\|^2 \cdot \|\Lambda^{(l)}\|^2$. Thus, the iterative procedure is equivalent to the well-known *power method* of Hotelling (1933); after normalizations, the sequence $F^{(0)}, F^{(1)}, \ldots$ will converge to the eigenvector associated with the largest eigenvalue of XX', as in the PCA estimator of Bai (2003). Therefore, the IQR algorithm, and its corresponding QFA estimator, can be viewed as an extension of PCA to QFM using QR tools.

Similar algorithms have been proposed in the machine learning literature to reduce the dimensions for binary data, where the check function is replaced by some smooth nonlinear link functions, e.g., Collins et al. (2002). However, unlike PCA, whether such methods guarantee finding the global minimum remains an open question. Nonetheless, in all of our Monte Carlo simulations we found that the QFA estimators of the factors using the IQR algorithm always converge to the space of the true factors, which is somewhat reassuring in this respect.

To prove the consistency of the QFA estimator $\hat{\theta}$, we make the following assumptions:

Assumption 1. *(i) $\mathcal{A}$ and $\mathcal{F}$ are compact sets and $\theta_0 \in \Theta^M$. In particular, $N^{-1}\sum_{i=1}^{N} \lambda_{0i}\lambda_{0i}' = diag(\sigma_{N1}, \ldots, \sigma_{Nr})$ with $\sigma_{N1} \geq \sigma_{N2} \cdots \geq \sigma_{Nr}$, and $\sigma_{Nj} \to \sigma_j$ as $N \to \infty$ for $j = 1, \ldots, r$ with $\infty > \sigma_1 > \sigma_2 \cdots > \sigma_r > 0$.*
(ii) Let f_{it} denote the density function of u_{it} given $\{f_{0t}\}$. There exists $\underline{\mathsf{f}} > 0$ such that for any compact set $C \subset \mathbb{R}$ and any $u \in C$, $\mathsf{f}_{it}(u) \geq \underline{\mathsf{f}}$ for all i, t.
(iii) Given $\{f_{0t}\}$, u_{it} is independent of u_{js} for any $i \neq j$ or $s \neq t$.

Write $\hat{\Lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_N)'$, $\Lambda_0 = (\lambda_{01}, \ldots, \lambda_{0N})'$, $\hat{F} = (\hat{f}_1, \ldots, \hat{f}_T)'$, $F_0 = (f_{01}, \ldots, f_{0T})'$, and let $L_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. The following theorem provides the average rate of convergence of $\hat{\Lambda}$ and $\hat{F}$.

Theorem 1. *Under Assumption 1, as $N, T \to \infty$, we have*

$$\|\hat{\Lambda} - \Lambda_0\|/\sqrt{N} = O_P(1/L_{NT}) \quad \text{and} \quad \|\hat{F} - F_0\|/\sqrt{T} = O_P(1/L_{NT}).$$

Remark 1.1: Since our proof strategy is substantially different from the one in Bai and Ng (2002), we briefly sketch the main ideas underlying our proof here. To facilitate the discussion, for any $\theta_a, \theta_b \in \Theta^M$ define the semimetric d by:

$$d(\theta_a, \theta_b) = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\lambda_{ai}' f_{at} - \lambda_{bi}' f_{bt})^2} = \frac{1}{\sqrt{NT}}\left\|\Lambda_a F_a' - \Lambda_b F_b'\right\|,$$

and let

$$\bar{\mathbb{M}}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[\rho_\tau(X_{it} - \lambda_i' f_t)].$$

The semimetric d plays an important role in our asymptotic analysis. We first show that $d(\hat{\theta}, \theta_0) = o_P(1)$. Next, it can be shown that:

$$\bar{\mathbb{M}}_{NT}(\hat{\theta}) - \bar{\mathbb{M}}_{NT}(\theta_0) \gtrsim d^2(\hat{\theta}, \theta_0), \tag{4}$$

and that for sufficiently small $\delta > 0$,

$$\mathbb{E}\left[\sup_{\theta\in\Theta^M(\delta)} \left|\mathbb{M}_{NT}(\theta) - \bar{\mathbb{M}}_{NT}(\theta) - \mathbb{M}_{NT}(\theta_0) + \bar{\mathbb{M}}_{NT}(\theta_0)\right|\right] \lesssim \frac{\delta}{L_{NT}}, \tag{5}$$

where $\Theta^M(\delta) = \{\theta \in \Theta^M : d(\theta, \theta_0) \leq \delta\}$. Intuitively, the above two inequalities and $d(\hat{\theta}, \theta_0) = o_P(1)$ imply that $d^2(\hat{\theta}, \theta_0) \lesssim d(\hat{\theta}, \theta_0)/L_{NT}$, or $d(\hat{\theta}, \theta_0) \lesssim L_{NT}^{-1}$. Then, the desired results follow from the fact that $\|\hat{\Lambda} - \Lambda_0\|/\sqrt{N} + \|\hat{F} - F_0\|/\sqrt{T} \lesssim d(\hat{\theta}, \theta_0)$.

Inequality (4) follows easily from a Taylor expansion of $\bar{\mathbb{M}}_{NT}(\hat{\theta})$ around θ_0 and Assumption 1(ii). It is worth stressing that the proof of (5) requires the chaining argument which is commonly used in the theory of empirical processes. In particular, using Hoeffding's inequality and the fact that $|\rho_\tau(u) - \rho_\tau(v)| \leq 2|u - v|$, it can be shown that, for any given $\theta_a, \theta_b \in \Theta^M$,

$$P\left[\sqrt{NT}\left|\mathbb{M}_{NT}(\theta_a) - \bar{\mathbb{M}}_{NT}(\theta_a) - \mathbb{M}_{NT}(\theta_b) + \bar{\mathbb{M}}_{NT}(\theta_b)\right| \geq c\right] \leq e^{-\frac{c^2}{Kd^2(\theta_a,\theta_b)}} \tag{6}$$

for some constant K. Then, along the lines of Theorem 2.2.4 of Van der Vaart and Wellner (1996), it follows that the left-hand side of (5) is bounded by $\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^M(\delta))}d\epsilon/\sqrt{NT}$. Finally, for sufficiently small δ, the semimetric d is shown to be equivalent to the Euclidean norm in $\mathbb{R}^M$, thus we can prove that $\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^M(\delta))}d\epsilon \lesssim \delta\sqrt{M}$, from which inequality (5) follows.

Remark 1.2: Compared to Bai and Ng (2002), notice that we do not require any moment of u_{it} to be finite. Thus, for the canonical factor models (e.g., Example 1) where the idiosyncratic errors have median equal to zero, our estimator for the case $\tau = 0.5$ can be interpreted as a least absolute deviation (LAD) estimator which is robust to heavy tails and outliers. In Section 5, we will illustrate the robustness of the LAD estimator, relative to the PCA estimator, by means of Monte Carlo simulations.

Remark 1.3: If the true parameters do not satisfy the normalizations (3), they can still be in the space Θ^M after some normalizations. Let H_{NT} be a $r \times r$ invertible matrix and define $\bar{f}_{0t} = H_{NT}' f_{0t}$, $\bar{\lambda}_{0i} = (H_{NT})^{-1}\lambda_{0i}$. Note that $\lambda_{0i}' f_{0t} = \bar{\lambda}_{0i}'\bar{f}_{0t}$. For $\{\bar{f}_{0t}\}$ and $\{\bar{\lambda}_{0i}\}$ to satisfy the normalizations (3), we require:

$$\frac{1}{T}\sum_{t=1}^{T}\bar{f}_{0t}\bar{f}_{0t}' = H_{NT}'\Sigma_{T,F}H_{NT} = \mathbb{I}_r \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^{N}\bar{\lambda}_{0i}\bar{\lambda}_{0i}' = (H_{NT})^{-1}\Sigma_{N,\Lambda}(H_{NT}')^{-1} = \mathbb{D}_N$$

where $\Sigma_{T,F} = T^{-1}\sum_{t=1}^{T} f_{0t}f_{0t}'$, $\Sigma_{N,\Lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda_{0i}\lambda_{0i}'$, and $\mathbb{D}_N$ is a diagonal matrix with non-increasing diagonal elements. The above equalities imply that:

$$\Sigma_{T,F}^{1/2}\Sigma_{N,\Lambda}\Sigma_{T,F}^{1/2}\cdot\Sigma_{T,F}^{1/2}H_{NT} = \Sigma_{T,F}^{1/2}H_{NT}\cdot\mathbb{D}_N.$$

Thus, the rotation matrix H_{NT} can be chosen as $\Sigma_{T,F}^{-1/2}\Gamma_{NT}$, where Γ_{NT} is the matrix of eigenvectors of $\Sigma_{T,F}^{1/2}\Sigma_{N,\Lambda}\Sigma_{T,F}^{1/2}$. As a result, Theorem 1 can be stated as follows:

$$\|\hat{\Lambda} - \Lambda_0(H_{NT}')^{-1}\|/\sqrt{N} = O_P(1/L_{NT}) \quad \text{and} \quad \|\hat{F} - F_0H_{NT}\|/\sqrt{T} = O_P(1/L_{NT}).$$

Note that the rotation matrix H_{NT} is slightly different from the rotation matrix of Bai (2003), but they converge to the same limit as $N, T \to \infty$ (see Remark 4.3 below).

Remark 1.4: Compared to Bai and Ng (2002), our Assumption 1(iii) is admittedly strong. However, note that this assumption is made conditional on $\{f_{0t}\}$, so cross-sectional dependence of u_{it} due to the common factors is still allowed for. Moreover, the independence assumption is only used to establish the sub-Gaussian inequality (6). Thus, Assumption 1(iii) can be relaxed as long as the sub-Gaussian inequality holds.[6]

3.2 Selecting the Number of Factors

In the previous section, we assumed the number of quantile-dependent factors $r(\tau)$ to be known at each τ. In this subsection we propose two different procedures to select the correct number of factors at each quantile with probability approaching one. The first one selects the model by rank minimization while the second one uses information criteria (IC). As before, the dependence of the quantile-dependent objects on τ, including $r(\tau)$, is ignored in the sequel.

3.2.1 Model Selection by Rank Minimization

Let k be a positive integer larger than r, and $\mathcal{A}^k$ and $\mathcal{F}^k$ be compact subsets of $\mathbb{R}^k$. In particular, let us assume that $[\lambda_{0i}' \quad \mathbf{0}_{1\times(k-r)}]' \in \mathcal{A}^k$ for all i. Let $\lambda_i^k, f_t^k \in \mathbb{R}^k$ for all i,t and write $\theta^k = (\lambda_1^{k\prime},\ldots,\lambda_N^{k\prime}, f_1^{k\prime},\ldots,f_T^{k\prime})'$, $\Lambda^k = (\lambda_1^k,\ldots,\lambda_N^k)'$, $F^k = (f_1^k,\ldots,f_T^k)'$. Consider the following

[6] See van de Geer (2002) for the properties of Hoeffding inequalities for martingales.

normalizations:

$$\frac{1}{T}\sum_{t=1}^{T} f_t^k f_t^{k\prime} = \mathbb{I}_k, \quad \frac{1}{N}\sum_{i=1}^{N} \lambda_i^k \lambda_i^{k\prime} \text{ is diagonal with non-increasing diagonal elements.} \tag{7}$$

Define $\Theta^k = \{\theta^k : \lambda_i^k \in \mathcal{A}^k, f_t^k \in \mathcal{F}^k, \text{ and } \lambda_i^k, f_t^k \text{ satisfy (7)}\}$, and

$$\hat{\theta}^k = (\hat{\lambda}_1^{k\prime}, \ldots, \hat{\lambda}_N^{k\prime}, \hat{f}_1^{k\prime}, \ldots, \hat{f}_T^{k\prime})' = \arg\min_{\theta^k \in \Theta^k} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \rho_\tau(X_{it} - \lambda_i^{k\prime} f_t^k).$$

Moreover, define $\hat{\Lambda}^k = (\hat{\lambda}_1^k, \ldots, \hat{\lambda}_N^k)'$ and write

$$(\hat{\Lambda}^k)'\hat{\Lambda}^k/N = \text{diag}\left(\hat{\sigma}_{N,1}^k, \ldots, \hat{\sigma}_{N,k}^k\right).$$

The first estimator of the number of factors r is defined as:

$$\hat{r}_{\text{rank}} = \sum_{j=1}^{k} \mathbf{1}\{\hat{\sigma}_{N,j}^k > P_{NT}\},$$

where P_{NT} is a sequence that goes to 0 as $N, T \to \infty$. In other words, $\hat{r}_{\text{rank}}$ is equal to the number of diagonal elements of $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$ that are larger than the threshold P_{NT}. We call $\hat{r}_{\text{rank}}$ the *rank-minimization estimator* because, as discussed below in Remark 2.1, it can be interpreted as a rank estimator of $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$.

It can be shown that:

Theorem 2. *Under Assumption 1, $P[\hat{r}_{rank} = r] \to 1$ as $N, T \to \infty$ if $k > r$, $P_{NT} \to 0$ and $P_{NT}L_{NT}^2 \to \infty$.*

Remark 2.1: In the proof of Theorem 2, we show that for $k > r$, it holds that

$$\left\|\hat{F}^{k,r} - F_0\right\|/\sqrt{T} = O_P(1/L_{NT}) \quad \text{and} \quad \left\|\hat{\Lambda}^k - \Lambda_0^*\right\|/\sqrt{N} = O_P(1/L_{NT}),$$

where $\hat{F}^{k,r}$ is the first r columns of $\hat{F}^k$ and $\Lambda_0^* = [\Lambda_0, \mathbf{0}_{N\times(k-r)}]$. It then follows from Assumption 1 that $\hat{\sigma}_{N,j}^k \xrightarrow{p} \sigma_j > 0$ for $j = 1, \ldots, r$ and $\hat{\sigma}_{N,j}^k = N^{-1}\sum_{i=1}^{N}\left(\hat{\lambda}_{ji}^k\right)^2 = O_P(1/L_{NT}^2)$ for $j = r+1, \ldots, k$. Thus, the first r diagonal components of $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$ converge in probability to positive constants while the remaining diagonal components are all $O_P(1/L_{NT}^2)$. In other words, $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$ converges to a matrix with rank r, and P_{NT} can be viewed as a cutoff value to choose the asymptotic rank of $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$.

3.2.2 Model Selection by Information Criteria

The second estimator of r is similar to the IC-based estimator of Bai and Ng (2002). Let l denote a positive integer smaller than or equal to k, and let $\mathcal{A}^l$ and $\mathcal{F}^l$ be compact subsets of $\mathbb{R}^l$. In particular, for $l > r$, assume that $[\lambda_{0i}' \quad \mathbf{0}_{1\times(l-r)}]' \in \mathcal{A}^l$ for all i. Moreover, we can define $\Theta^l, \hat{\theta}^l, \hat{f}_t^l, \hat{\lambda}_i^l, \hat{F}^l$ and $\hat{\Lambda}^l$ in a similar fashion.

Define the IC-based estimator of r as follows:

$$\hat{r}_{\mathrm{IC}} = \underset{1\leq l\leq k}{\arg\min}\left[\mathbb{M}_{NT}(\hat{\theta}^l) + l \cdot P_{NT}\right].$$

We can show that:

Theorem 3. *Suppose Assumption 1 holds, and assume that there exists $\bar{\mathsf{f}} > 0$ such that for any compact set $C \subset \mathbb{R}$ and any $u \in C$, $\mathsf{f}_{it}(u) \leq \bar{\mathsf{f}}$ for all i, t. Then $P[\hat{r}_{IC} = r] \to 1$ as $N, T \to \infty$ if $k > r$, $P_{NT} \to 0$ and $P_{NT} L_{NT}^2 \to \infty$.*

Remark 3.1: A similar result is also obtained by AB (2018), but the difference with ours is that we only need the density function of the idiosyncratic errors to be uniformly bounded above and below, while AB (2018) requires all the moments of the errors to be bounded. This difference is crucial since the robustness of our estimators against heavy tails and outliers becomes their main advantage relative to PCA estimators. The reason why we can obtain the same result here with less restrictions is that our proof is based on the innovative argument discussed in Remark 1.1 and the average convergence rate of the estimators, while the proof of AB (2018) depends on the uniform convergence rate of the estimators.

Remark 3.2: Note that, for AFM, the rank-minimization estimator and the IC-based estimator of r are equivalent. To see this, let X denote the $T \times N$ matrix of observed variables, and let $\check{F}^l, \check{\Lambda}^l$ denote the matrices of PCA estimators of Bai and Ng (2002) when the estimated number of factors is l. Then Bai and Ng (2002)'s estimator of r can be written as:

$$\hat{r} = \underset{1\leq l\leq k}{\arg\min}\,\hat{S}(l) \quad \text{where} \quad \hat{S}(l) = (NT)^{-1}\left\|X - \check{F}^l\check{\Lambda}^{l\prime}\right\|^2 + l \cdot P_{NT},$$

$k > r$, and P_{NT} is defined as in Theorem 2 above. Since $\check{F}^l/\sqrt{T}$ are the l eigenvectors of $XX'/(NT)$ associated with the largest l eigenvalues and $\check{\Lambda}^l = X'\check{F}^l/T$, we have that:

$$(NT)^{-1}\left\|X - \check{F}^l\check{\Lambda}^{l\prime}\right\|^2 = \mathrm{Tr}[XX'/(NT)] - \mathrm{Tr}\left[\check{F}^{l\prime}/\sqrt{T}(XX'/(NT))\check{F}^l/\sqrt{T}\right] = \sum_{j=l+1}^{T}\rho_j\left(XX'/(NT)\right).$$

Therefore, $\hat{S}(l) - \hat{S}(l-1) = P_{NT} - \rho_l\left(XX'/(NT)\right)$, and $\hat{S}(l)$ is minimized at $\hat{r}$ if

$$\rho_{\hat{r}}\left(XX'/(NT)\right) > P_{NT} \quad \text{and} \quad \rho_{\hat{r}+1}\left(XX'/(NT)\right) \leq P_{NT}.$$

That is, $\hat{r}$ is chosen as the number of eigenvalues of $XX'/(NT)$ that are larger than P_{NT}. Further, let $\rho_1(X) \geq \ldots \geq \rho_k(X)$ be the k largest eigenvalues of $XX'/(NT)$, then it is easy to see that:

$$\operatorname{diag}\left(\rho_1(X), \ldots, \rho_k(X)\right) = \check{F}^{k'}/\sqrt{T}(XX'/(NT))\check{F}^k/\sqrt{T} = \check{\Lambda}^{k'}\check{\Lambda}^k/N.$$

Therefore, Bai and Ng (2002)'s estimator of r is equivalent to the number of diagonal elements in $\check{\Lambda}^{k'}\check{\Lambda}^k/N$ that are larger than P_{NT} — which is equivalent to the rank-minimization estimator that we defined above. However, due to the differences of the object functions, such equivalence does not exist in QFM.

Remark 3.3: The choice of P_{NT} for $\hat{r}_{\text{rank}}$ and $\hat{r}_{\text{IC}}$ can be different in practice. In particular, it can differ from those penalties used by Bai and Ng (2002). AB (2018) choose

$$P_{NT} = \log\left(\frac{NT}{N+T}\right) \cdot \frac{N+T}{NT}$$

for $\hat{r}_{\text{IC}}$, similar to IC_{p1} of Bai and Ng (2002). However, as shown in AB's (2018) simulation results, this choice does not perform very well even for N, T as large as 300.

Remark 3.4: Even though $\hat{r}_{\text{rank}}$ and $\hat{r}_{\text{IC}}$ are both consistent estimators of r, the computational cost of $\hat{r}_{\text{rank}}$ is much lower than that of $\hat{r}_{\text{IC}}$, because for $\hat{r}_{\text{rank}}$ we only estimate the model once, while for $\hat{r}_{\text{IC}}$ we need to estimate the model k times. Thus, in the simulations we will focus on $\hat{r}_{\text{rank}}$, and we refer to AB (2018) for the corresponding simulation results of $\hat{r}_{\text{IC}}$. We find that the choice

$$P_{NT} = \hat{\sigma}^k_{N,1} \cdot \left(\frac{1}{L^2_{NT}}\right)^{1/3}$$

for $\hat{r}_{\text{rank}}$ works fairly well as long as $\min\{N, T\}$ is 100. This is also the value used in all of our simulations and applications.

3.3 Discriminating between AFM and QFM

The asymptotic results above guarantee that the QFA approach for QFM is not simply overfitting the data by estimating more spurious factors. Hence, it provides a sensible alternative procedure to PCA for estimation of factor structures. As a result, a relevant issue in practice is whether differences between the estimated number of QFA and PCA factors can help discriminating between AFM and QFM structures.

Before addressing this issue, however, it is worth highlighting that such a comparison does not provide a formal test of AFM vs. QFM. In effect, using the analogy of OLS regressions vs. QR, finding that the QR estimated coefficients vary across the conditional quantiles of the dependent variable does not imply that the OLS results are invalid. As it is well known, this is because OLS estimation focuses on the average response of the dependent variable to a change in an explanatory variable, whereas QR looks at how such a response varies throughout of the distribution of the dependent variable. Thus, since these two estimation procedures have different goals (modelling conditional means and conditional quantiles), the only valid claim one can make is that is that QR provides larger information insofar as the estimates differ across quantiles.

In the FM model literature, AFM is not tested in a formal way. It is instead selected by some consistent selection criteria: there is an AFM insofar $0 < r \ll N$, and then the chosen factors and loadings are estimated by PCA (or other similar estimation procedures). Following the same reasoning, selection between AFM and QFM relies on the comparison of the number of estimated factors by PCA and QFA, which for convenience we label $\hat{r}_{PCA}$ and $\hat{r}_{QFA}(\tau)$, respectively, in the sequel.

Then, according to the number of factors estimated by each procedure, the following two cases could be distinguished:

(I) If $\hat{r}_{QFA}(\tau) > \hat{r}_{PCA}$ for some τs, this ensures the existence of extra factors, so that the QFA estimation approach is needed to extract them. Example 3 in subsection 2.2 above provides a simple illustration of such a case. The IC-based selection criteria of Bai and Ng (2002) will choose r_1 PCA factors, while our two consistent selection criteria will select $r_1 + r_2$ QFA factors (except at $\tau = 0.5$ where they will choose r_1). Similar arguments apply to Examples 4 and 5 above.[7]

(II) If $\hat{r}_{QFA}(\tau) \leq \hat{r}_{PCA}$ for all τs, there could still exist some extra factors which differ from the mean-shifting factors detected by PCA. This could happen if the loadings of some extra factors are zero or small for certain τs. In such instances, QFA will find it difficult to detect them in finite samples, resembling the issues raised by Onatski (2012) about the role of weak factors in AFM. A potential illustration, which is not listed in subsection 2.2 above, could be the following QFM structure: $X_{it} = \lambda_{1i}(\tau)f_{1t} + \lambda_{2i}(\tau)f_{2t} + u_{it}(\tau)$, where f_{1t} is a mean-shifting factor, and f_{2t} only affects the upper and lower quantiles but not the mean of X_{it}, i.e., $\lambda_{2i}(\tau) = 0$ for $\tau \in [\epsilon, 1 - \epsilon]$. If $\lambda_{1i}(\tau)$ is close to zero for $\tau \in (0, \epsilon) \cup (1 - \epsilon, 1)$, so that f_{1t} is a weak factor in those parts of the distribution where f_{2t} hits, then QFA will only capture f_{2t} but not f_{1t} at the upper and lower quantiles, while PCA will only capture f_{1t} but not f_{2t}. In this example,

[7] Within this category, one could also include the standard FM in Example 1, since the number of QFA factors at all τs (except $\tau = 0.5$) would exceed the number of PCA factors by exactly one factor, namely, a unit vector. Thus, QFA will easily detect this case because one of the two selected factors will be highly correlated with the PCA factor, while the other factor will be constant over time.

it becomes evident that, despite yielding the same number of PCA and QFA factors at all τs, the factors are not the same. That would not hold in Example 2 of subsection 2.2, where PCA and QFA select the same number of factors (equal to 1) and the selected factor by each estimation method happens to be identical (f_{1t}). Thus, whenever the difference between $\hat{r}_{QFA}$ and $\hat{r}_{PCA}$ falls into this range, our suggestion to check if PCA captures all the factors in the QFM representation (like in Example 2) relies on computing correlations between the estimated QFA factors at different τs and the PCA factors. If these correlations are high, this will be an indication that PCA extracts all the relevant factors in the QFM representation, while if they are low for some τs, this will be signaling that PCA fails to do so.

This discrimination strategy between AFM and QFM will be subject to further discussion in Section 6 below when we apply it to interpret results in our empirical applications.

4 Estimators Based on Smoothed Quantile Regressions

The asymptotic distribution of the QFA estimator $\hat{\theta}$ is difficult to derive due to the non-smoothness of the check function and the problem of incidental parameters. As in the asymptotic analysis of standard QR, one can expand the expected score function (which is smooth and continuously differentiable) and obtain a stochastic expansion for $\hat{\lambda}_i - \lambda_{0i}$; yet the following term appears in the expansion:

$$\frac{1}{T}\sum_{t=1}^{T}\left\{\left(\mathbf{1}\{X_{it} \leq \hat{\lambda}_i'\hat{f}_t\} - \mathbb{E}[\mathbf{1}\{X_{it} \leq \hat{\lambda}_i'\hat{f}_t\}]\right)\hat{f}_t - \left(\mathbf{1}\{X_{it} \leq \lambda_{0i}'f_{0t}\} - \tau\right)f_{0t}\right\}. \tag{8}$$

AB (2018) claim that the above term is $o_P(1/T^{1/2})$, based on the results that $\max_{i\leq N}\|\hat{\lambda}_i - \lambda_{0i}\| = o_P(1)$ and $\max_{t\leq T}\|\hat{f}_t - f_{0t}\| = o_P(1)$. However, we suspect that this claim may not hold. To see this, let and $\check{\lambda}_i$ and $\check{f}_t$ be the PCA estimators in a AFM. In the stochastic expansion of $\check{\lambda}_i - \lambda_{0i}$, the analogous term to (8) happens to be:

$$\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}(\check{f}_t - f_{0t}),$$

where ϵ_{it} is the idiosyncratic error in the AFM. Note that, based on $\max_{t\leq T}\|\check{f}_t - f_{0t}\| = o_P(1)$, one can only show that:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}(\check{f}_t - f_{0t})\right\| \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2}\cdot\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\check{f}_t - f_{0t})\|^2} = o_P(1).$$

Instead, one has to use the stochastic expansion of $\check{f}_t - f_{0t}$ to show that $T^{-1}\sum_{t=1}^{T}\epsilon_{it}(\check{f}_t - f_{0t}) = 1/L_{NT}^2$ (see the proof of Lemma B.1 of Bai 2003). Likewise, to show that (8) is $o_P(1/T^{1/2})$, and therefore that this term does not affect the asymptotic distribution of $\hat{\lambda}_i$, establishing the convergence rate of $\hat{f}_t - f_{0t}$ is not enough. As a result, the stochastic expansion of $\hat{f}_t - f_{0t}$ is needed. However, due the non-smoothness of the indicator functions, it is not clear how to explore the stochastic expansion of $\hat{f}_t - f_{0t}$ in (8).

To overcome the problem discussed above, we proceed to define a new estimator of θ_0, denoted as $\tilde{\theta}$, based on the following smoothed quantile regressions (SQR):

$$\tilde{\theta} = (\tilde{\lambda}_1', \ldots, \tilde{\lambda}_N', \tilde{f}_1', \ldots, \tilde{f}_T')' = \arg\min_{\theta \in \Theta^M} \mathbb{S}_{NT}(\theta),$$

where

$$\mathbb{S}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\tau - K\left(\frac{X_{it} - \lambda_i' f_t}{h}\right)\right](X_{it} - \lambda_i' f_t),$$

$K(z) = 1 - \int_{-1}^{z} k(z)dz$, $k(z)$ is a continuous function with support $[-1, 1]$, and h is a bandwidth parameter that goes to 0 as N, T diverge.

Define

$$\Phi_i = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathsf{f}_{it}(0) f_{0t} f_{0t}' \quad \text{and} \quad \Psi_t = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathsf{f}_{it}(0)\lambda_{0i}\lambda_{0i}'$$

for all i, t. We impose the following assumptions:

Assumption 2. *(i) $\Phi_i > 0$ and $\Psi_t > 0$ for all i, t.*
(ii) λ_{0i} is an interior point of $\mathcal{A}$ and f_{0t} is an interior point of $\mathcal{F}$ for all i, t.
(iii) $k(z)$ is symmetric around 0 and twice continuously differentiable. For $m \geq 8$, $\int_{-1}^{1} k(z)dz = 1$, $\int_{-1}^{1} z^j k(z)dz = 0$ for $j = 1, \ldots, m-1$ and $\int_{-1}^{1} z^m k(z)dz \neq 0$.
(iv) f_{it} is $m+2$ times continuously differentiable. Let $\mathsf{f}_{it}^{(j)}(u) = (\partial/\partial u)^j \mathsf{f}_{it}(u)$ for $j = 1, \ldots, m+2$. There exists $-\infty < \underline{l} < \bar{l} < \infty$, such that for any compact set $C \subset \mathbb{R}$ and any $u \in C$, we have $\underline{l} \leq \mathsf{f}_{it}^{(j)}(u) \leq \bar{l}$ and $\underline{\mathsf{f}} \leq \mathsf{f}_{it}(u) \leq \bar{l}$ for $j = 1, \ldots, m+2$ and for all i, t.
(v) As $N, T \to \infty$, $N \propto T$, $h \propto T^{-c}$ and $m^{-1} < c < 1/6$.

Then, we can show that:

Theorem 4. *Under Assumptions 1 and 2,*

$$\sqrt{T}(\tilde{\lambda}_i - \lambda_{0i}) \xrightarrow{d} \mathcal{N}(0, \tau(1-\tau)\Phi_i^{-2}) \quad \textit{and} \quad \sqrt{N}(\tilde{f}_t - f_{0t}) \xrightarrow{d} \mathcal{N}(0, \tau(1-\tau)\Psi_t^{-1}\Sigma_\Lambda\Psi_t^{-1})$$

for each i and t, where $\Sigma_\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$.

Remark 4.1: Similar to the proof of Theorem 1, we can show that

$$\|\tilde{\Lambda} - \Lambda_0\|/\sqrt{N} = O_P(1/L_{NT}) + O_P(h^{m/2}) \quad \text{and} \quad \|\tilde{F} - F_0\|/\sqrt{T} = O_P(1/L_{NT}) + O_P(h^{m/2}),$$

where the extra $O_P(h^{m/2})$ term is due the approximation bias of the smoothed check function. However, Assumption 2(v) implies that $1/L_{NT} >> h^{m/2}$, and then it follows that average convergence rates of $\tilde{\Lambda}$ and $\tilde{F}$ are both L_{NT}.

Remark 4.2: Similar to Theorems 1 and 2 of Bai (2003), we show that the new estimator is free of incidental-parameter biases. That is, the asymptotic distribution of $\tilde{\lambda}_i$ is the same as if we would observe $\{f_{0t}\}$, and likewise the asymptotic distribution of $\tilde{f}_t$ is the same as if $\{\lambda_{0i}\}$ were observed. The proof of this result is not trivial. To see why this is the case, first define $\varrho(u) = [\tau - K(u/h)]u$ and $\mathbb{S}_{i,T}(\lambda, F) = T^{-1}\sum_{t=1}^{T} \varrho(X_{it} - \lambda' f_t)$, then we can write $\tilde{\lambda}_i = \arg\min_{\lambda\in\mathcal{A}} \mathbb{S}_{i,T}(\lambda, \tilde{F})$. Expanding $\partial\mathbb{S}_{i,T}(\tilde{\lambda}_i, \tilde{F})/\partial\lambda$ around (λ_{0i}, F_0) yields

$$\begin{aligned}\left(\frac{1}{T}\sum_{t=1}^{T} \varrho^{(2)}(u_{it}) f_{0t} f_{0t}'\right)(\tilde{\lambda}_i - \lambda_{0i}) \approx \frac{1}{T}\sum_{t=1}^{T} \varrho^{(1)}(u_{it}) f_{0t} + \frac{1}{T}\sum_{t=1}^{T} \rho^{(1)}(u_{it})(\tilde{f}_t - f_{0t}) \\ - \frac{1}{T}\sum_{t=1}^{T} \rho^{(2)}(u_{it}) f_{0t}\lambda_{0i}'(\tilde{f}_t - f_{0t}), \quad (9)\end{aligned}$$

where $\varrho^{(j)}(u) = (\partial/\partial u)^j \varrho(u)$. The key step is to show that the last two terms on the right-hand side of the above equation are $o_P(1/\sqrt{T})$. This is relatively easier for the PCA estimator of Bai (2003), since $(\tilde{f}_t - f_{0t})$ has an analytical form (e.g., equation A.1 of Bai 2003). In our case, we would need a similar expansion as (9) to obtain an approximate expression for $(\tilde{f}_t - f_{0t})$, but this expression depends on $(\tilde{\lambda}_i - \lambda_{0i})$ due to the nature of factor models. Similar to Chen et al. (2018), this problem can be partly solved by showing that the expected Hessian matrix is asymptotically block-diagonal (see Lemma 11 in the Appendix). However, the proof of Chen et al. (2018) is only applicable to a special infeasible normalization, namely $\sum_{i=1}^{N} \lambda_{0i}\lambda_i = \sum_{t=1}^{T} f_{0t} f_t'$, while our proof of Lemma 11 allows for normalization (3) and can be generalized to any of the other normalizations considered by Bai and Ng (2013) that uniquely pin down the rotation matrix.

Remark 4.3: As discussed in Remark 1.3, if the true parameters do not satisfy the normalizations (3), the results of Theorem 3 can be stated as

$$\sqrt{T}\left(\tilde{\lambda}_i - H_{NT}^{-1}\lambda_{0i}\right) \xrightarrow{d} \mathcal{N}\left(0, \tau(1-\tau)H^{-1}\Phi_i^{-1}\Sigma_F\Phi_i^{-1}(H'^{-1})\right),$$

$$\sqrt{N}\left(\tilde{f}_t - H_{NT}' f_{0t}\right) \xrightarrow{d} \mathcal{N}\left(0, \tau(1-\tau)H'\Psi_t^{-1}\Sigma_\Lambda\Psi_t^{-1}H\right),$$

where $\Sigma_F = \lim_{T\to\infty}\Sigma_{T,F}$, $\Sigma_\Lambda = \lim_{N\to\infty}\Sigma_{N,\Lambda}$, $H = \Sigma_F^{-1/2}\Gamma$, and Γ is the matrix of eigenvectors of $\Sigma_F^{1/2}\Sigma_\Lambda\Sigma_F^{1/2}$.

Remark 4.4: A restrictive DGP within class (1) would be a QFM where the PCA factors coincide with the quantile factors and only the factor loadings are quantile dependent. The representation for such restricted subset of QFM is as follows:

$$X_{it} = \lambda_i'(\tau) f_t + u_{it}(\tau), \text{ for } \tau \in (0,1). \tag{10}$$

As a result, the main objects of interest are the common factors and the quantile-varying loadings. Notice that, if the factors f_t were to be observed, using standard QR of X_{it} on f_t would lead to consistent and asymptotically normally distributed estimators of $\lambda_i(\tau)$ for each i and $\tau \in \mathcal{T}$. However, since f_t are not observable, a feasible two-stage approach is to first estimate the factors by PCA, denoted as $\hat{f}_t^{PCA}$, and next run QR of X_{it} on $\hat{f}_t^{PCA}$ to obtain estimates of $\lambda_i(\tau)$ as follows:

$$\hat{\lambda}_i(\tau) = \arg\min_{\lambda} T^{-1} \sum_{t=1}^{T} \rho_\tau(X_{it} - \lambda' \hat{f}_t^{PCA}). \tag{11}$$

As explained in Chen et al. (2017), unlike the QFA estimators, this two-stage procedure requires moments of the idiosyncratic term u_{it} to be bounded in order to apply PCA in the first stage (see Remark 1.2). However, an interesting result (see Chen et al. 2017, Theorem 2) is that the standard conditions on the relative asymptotics of N and T allowing for the estimated factors to be treated as known do not hold when applying this two-stage estimation approach. In effect, while these conditions are $T^{1/2}/N \to 0$ for linear factor-augmented regressions (see Bai and Ng 2006) and $T^{5/8}/N \to 0$ for nonlinear factor-augmented regressions (Bai and Ng 2008a), lack of smoothness in the object (check) function at the second stage requires the stronger condition $T^{5/4}/N \to 0$. Moreover, Theorem 3 in Chen et al. (2017) shows how to run inference on the quantile-varying loadings (e.g., testing the null that they are constant across all quantiles or a subset of them).

5 Finite Sample Simulations

In this section we report the results from several Monte Carlo simulations regarding the performance of our proposed QFM methodology in finite samples. In particular, we focus on three relevant issues: (i) how well does our preferred estimator of the number of factors perform relative to other selection criteria when the distribution of the idiosyncratic error terms in an AFM exhibits heavy tails, (ii) how well do PCA and QFA estimate the true factors under the previous circumstances, and (iii) how robust is the QFA estimation procedure when the errors terms are serially and cross-sectionally correlated, instead of being independent.

5.1 Estimation of AFM: Heavy-tailed idiosyncratic error terms

As pointed out in Remark 1.2, our estimator for AFM at $\tau = 0.5$ can be viewed as a robust alternative to the PCA estimators that are commonly used in practice. This is because the consistency of our estimators does not require the moments of the idiosyncratic errors to exist. For the same reason, our estimator of the number of factors should also be more robust to outliers and heavy tails than the IC-based method of Bai and Ng (2002). In this subsection we confirm the above claims by means of simulations.

We consider the following DGP:

$$X_{it} = \sum_{j=1}^{3} \lambda_{ji} f_{jt} + u_{it},$$

where $f_{1t} = 0.2f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = 0.8f_{3,t-1} + \epsilon_{3t}$, $\lambda_{ji}, \epsilon_{jt}$ are all independent draws from $\mathcal{N}(0,1)$, and u_{it} are independent draws from the standard Cauchy distribution. We consider four estimators of the number of factors r: two estimators based on PC_{p1}, IC_{p1} of Bai and Ng (2002), the eigenvalue-ratio estimator of Ahn and Horenstein (2013) and our rank-minimization estimator discussed in subsection 3.2, having chosen

$$P_{NT} = \hat{\sigma}^{k}_{N,1} \cdot \left(\frac{1}{L^2_{NT}} \right)^{1/3}.$$

We set $k = 8$ for all four estimators, and consider $N, T \in \{50, 100, 200\}$.

Table 1 reports the following fractions:

$$[\text{proportion of } \hat{r} < 3 \text{ , proportion of } \hat{r} = 3 \text{ , proportion of } \hat{r} > 3 \text{ }]$$

for each estimator having run 1000 replications.

It becomes evident from the results in Table 1 that the IC-based estimators of Bai and Ng (2002) almost always overestimate the number factors, and that the eigenvalue-ratio estimator of Ahn and Horenstein (2013) tends to underestimate the number of factors but to a lesser extent than what the IC estimators overestimate them. By contrast, our rank-minimization estimator chooses accurately the right number of factors as long as $\min\{N, T\} \geq 100$.

Next, to compare the PCA and QFA estimators of the common factors in the previous DGP, we assume that $r = 3$ is known. We first get the PCA estimators $\hat{F}_{PCA}$, and then obtain the QFA estimator $\hat{F}_{QFA}$ using the IQR algorithm. Next, we regress each of the true factors on $\hat{F}_{PCA}$ and $\hat{F}_{QFA}$ separately, and report the average R^2 from 1000 replications in Table 2 as an indicator of how well the space of the true factors is spanned by the estimated factors. As shown in the first three columns of Table 2, while the PCA estimators are not very successful in capturing the true common factors, our QFA estimators approximate them very well, even when N, T are not too large.

As discussed earlier, the overall findings reported in Tables 1 and 2 are in line with our theoretical results. They provide strong evidence of the substantial gains that can be achieved by using QFA rather than PCA in those cases where the idiosyncratic error terms in AFM exhibit heavy tails and outliers.

5.2 Estimation of QFM: Heavy-tailed and non-independent error terms

In this subsection we consider the following DGP:

$$X_{it} = \lambda_{1i} f_{1t} + \lambda_{2i} f_{2t} + (\lambda_{3i} f_{3t}) \cdot e_{it},$$

where $f_{1t} = 0.8 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = |g_t|$, $\lambda_{1i}, \lambda_{2i}, \epsilon_{1t}, \epsilon_{2t}, g_t$ are all independent draws from $\mathcal{N}(0,1)$, and λ_{3i} are independent draws from $U[1,2]$. Following Bai and Ng (2002), the following specification for e_{it} is used:

$$e_{it} = \beta\ e_{i,t-1} + v_{it} + \rho \cdot \sum_{j=i-J, j\neq i}^{i+J} v_{jt},$$

where v_{it} are independent draws from $\mathcal{N}(0,1)$ except in the second case below. The autoregressive coefficient β captures the serial correlations of e_{it}, while the parameters ρ and J capture the cross-sectional correlations of e_{it}. We consider four cases:

Case 1: Independent errors: $\beta = 0$ and $\rho = 0$,
Case 2: Independent errors with heavy tails: $\beta = \rho = 0$, and $v_{it} \sim i.i.d$ Student(3).
Case 3: Serially Correlated Errors: $\beta = 0.2$ and $\rho = 0$.
Case 4: Serially and Cross-Sectionally Correlated Errors: $\beta = 0.2$ and $\rho = 0.2$, and $J = 3$.

For each of the previous cases and each $\tau \in \{0.25, 0.5, 0.75\}$, we first estimate $\hat{r}$ using our rank-minimization estimator, having set k and P_{NT} as described in the previous subsection. Second, we estimate $\hat{r}$ factors by means of the QFA estimation approach, which we denote as $\hat{F}^{\hat{r}}_{QFA}$. Finally, we regress each of the true factors on $\hat{F}^{\hat{r}}_{QFA}$ and calculate the R^2s. This procedure is repeated 1000 times and for each τ, we report the averages of $\hat{r}$ and the R^2s among these 1000 replications.

The results for Case 1 and Case 2 (where this time the heavy tails are captured by a Student(3) rather than by a Cauchy distribution) are reported in Tables 3 and 4, respectively, for $N, T \in \{50, 100, 200\}$. Notice that for $\tau = 0.25, 0.75$, we have $r(\tau) = 3$ while, for $\tau = 0.5$, we get $r(\tau) = 2$, since the factor f_{3t} does not affect the median of X_{it}. It can be observed that both our rank-minimization selection criterion and the QFA estimators perform very well in choosing the true number of QFA factors and in estimating them. It should be noticed that at $\tau = 0.25, 0.75$ the estimation of the scale factor f_{3t} is not as good as the mean factors f_{1t}, f_{2t} for small N and T. However, such differences vanish as N and T increase.

The results for Case 3 and Case 4 are in turn reported in Tables 5 and 6, respectively. It can be inspected that, even when the independence assumption is violated in these DGPs, the QFA estimation approach still performs satisfactorily. Thus, despite adopting independence in Assumption 1 (iii) for tractability in the proofs (see Remark 1.4), it seems that QFA estimation still works properly when the errors terms are allowed to exhibit mild serial and cross-sectional correlations.

6 Empirical Applications

In this section we illustrate the use of the QFA estimation approach in practice by considering three empirical applications that involve macroeconomic, financial, and climate change data:

1. The first dataset (SW for short) corresponds to an updated version of the popular panel of macroeconomic indicators which has been used by Stock and Watson to construct leading indicators for the US economy. This dataset can be downloaded from Mark Watson's website. SW consists of 167 quarterly macro-variables from 1959 to 2014 ($N = 167, T = 221$). These variable are transformed into stationary series before estimating the factors (see Stock and Watson 2016 for the details of this dataset).
2. The second dataset (Climate for short) consists of the annual changes of temperature from 338 stations from 1916 to 2016 ($N = 338, T = 100$) drawn from the Climate Research Unit (CRU) at the University of East Anglia, where information about global temperatures across different stations in the Northern and Southern Hemisphere is provided.
3. The third dataset (MF for short) contains the monthly returns of 2378 mutual funds from 2000 to 2014 ($N = 2378, T = 180$), obtained from the Center of Research for Security Prices (CRSP).

First, we set the number of PCA estimated factors in the SW dataset to be equal to 3 since this is the conventional number of factors found in the macroeconomic literature (typically capturing variability in TFP, monetary and fiscal variables). In contrast, for the Climate and MF datasets, which have been less explored in the AFM literature, we use the eigenvalue-ratio estimator of Ahn and Horenstein (2013) which selects 2 and 3 PCA factors, respectively; [8] next, we estimate the number of quantile-dependent factors using our rank-minimization estimator at $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$.

[8] We also applied the IC-based method of Bai and Ng (2002), but it was found that this selection procedure always chooses the maximum number of factors (8) for all the three datasets. For this reason, we only report the results of the eigenvalue-ratio estimator, whose finite-sample performance has been shown by Ahn and Horenstein (2013) to be more satisfactory than those of the IC-based methods and related selection rules.

The results of the previous exercise are reported in Table 7. Two different sets of findings stand out. On the one hand, there are two datasets where the estimated number of PCA and QFA factors across quantiles is quite similar or even identical. The first one is the SW dataset, where the estimated number of QFA factors using our rank-minimization estimator never differs from the estimated number of PCA factors (3) by more than one factor; for example, at $\tau = 0.10$ and 0.9, the chosen number of QFA factors is 2 while it is 4 at $\tau = 0.75$. In line with the discussion in subsection 3.3, our interpretation of these results is that some of the four selected quantile-varying factors may be relevant at $\tau = 0.75$, while they may be weak at the other two quantiles. The second one in this category is the MF dataset, where we find an even stronger degree of similarity between the number of QFA and PCA factors: for all considered τs, they are always identical (3).

On the other hand, the evidence for the Climate dataset is rather different. In effect, with the exception of two tails of the distribution ($\tau = 0.1$ and 0.9), where the estimated number of QFA factors equals the number of PCA factors (2), the selected number of QFA factors at the remaining quantiles (5 or 6) is much larger.

Thus, in line with the discussion in subsection 3.3, the previous findings for the Climate dataset strongly indicate that PCA fails to capture all relevant factors in the QFM representation, implying that the QFA estimation approach is required to extract them. Regarding the SW and MF datasets, it was also argued in subsection 3.3 that equality (or similarity) of the number of PCA and QFA factors at all considered τs does not necessarily imply that PCA captures all relevant factors. To check this, we examine the size of the correlation of each QFA factor at each τ with the set of estimated PCA factors. If these correlations are high, this would indicate that the QFA factors only capture the PCA factors, with no other extra factors being relevant. Conversely, if the correlations are low at some τ, this will indicate the presence of some extra factor at such a τ that PCA is unable to uncover.

Following this strategy, Table 8 shows the results of of comparing $\hat{F}_{FQA}$ with the PCA factors (denoted as $\hat{F}_{PCA}$).[9] For each τ, we regress each element of $\hat{F}_{QFA}$ on $\hat{F}_{PCA}$, and report the R^2s of these regressions. The main finding is that most of these R^2s are close to 1 (which is not surprising since mean-shifting factors affect most of the quantiles) but with a few noticeable exceptions: (i) the first QFA factor of SW at $\tau = 0.9$, (ii) the two QFA factors of Climate at $\tau = 0.1$ and 0.9, and (iii) the third QFA factor of MF at $\tau = 0.1$ and 0.25. These exceptions indicate that, besides the mean-shifting factors, the QFA estimation procedure is able to uncover other quantile-dependent factors which could provide extra information about the distributional characteristics of the data.

Finally, we further investigate the origins of these extra QFA factors so as to improve their

[9] As in Table 7, we estimate 3, 2 and 3 mean factors for SW, Climate and MF, respectively, whereas the number of QFA factors for each quantile τ also correspond to the figures displayed in Table 7.

interpretation. We do this by comparing them to the volatility factors obtained by the PCA-SQ procedure, denoted as $\hat{VF}_2$. The insight for this comparison can be provided by Example 3 above, where the extra QFA factors happen to be volatility factors and hence should be highly correlated. Furthermore, in a similar fashion, we also construct *skewness* factors and *kurtosis* factors by applying PCA to the third and fourth powers of the residuals obtained after removing the PCA factors from the data, which we denote as $\hat{VF}_3$ and $\hat{VF}_4$, respectively. Table 9 reports the R^2s of regressing $\hat{VF}_j$ on $\hat{F}_{QFA}$ for $j = 2, 3, 4$ at different τs. The results for the SW dataset are somewhat mixed. As can be observed in the first three rows of this Table, the explanatory power of the volatility, skewness and kurtosis factors over the QFA factors is fairly moderate. This evidence, together with the strong correlations between the PCA and QFA factors reported in Table 8, seems to point out that the three selected PCA factors play a dominant role in the QFM structure. Yet, in view of the slightly higher correlations (R^2s close to or above 0.6) of the QFA factors with $\hat{VF}_2$ at the lower and upper quantiles, one cannot rule out that extra factors related to volatility may still be relevant. By contrast, for Climate and MF, the evidence is much clearer: the skewness factor is highly correlated with the estimated QFA factors at most τs, whereas the volatility and kurtosis factors are not correlated at all with them. This finding points to the existence of common factors related to symmetry in the distribution of the variables included in these two datasets, which are properly captured by means of the QFA estimation procedure but omitted when applying PCA.

Interestingly, the evidence for the MF dataset is in line with the results by Andersen et al. (2018) who report the existence of tail factors in the distribution of asset returns which, for our specific dataset, we interpret as being closely related to changes in skewness. Likewise, the evidence for the Climate dataset is also in line with the results obtained by Gadea and Gonzalo (2019). Using the same dataset we use here (but different quantile techniques), these authors find that global warming over the last century seems to be due to a different behaviour in the lower tail than in the central and upper tails of the distribution of global temperatures. This finding points to a change in the skewness of such a distribution, in agreement with the nature of the extra QFA factors found for this dataset.

7 Conclusions

Approximate Factor Models (AFM) have become a leading methodology for the joint modelling of large number of economic time series with the big improvements in data collection and information technologies. This first generation of AFM was designed to reduce the dimensionality of big datasets by finding those common components which, by shifting the means of the observed variables with different intensities, are able to capture a large fraction of the data co-movements. However, one could envisage the existence of other common factors that do not (or not only) shift the means but also affect other distributional characteristics (volatility, higher moments, extreme values, etc.). This calls for a second generation of factor models.

Inspired by the generalization of linear regressions to quantile regressions (QR), this paper proposes Quantile Factor Models (QFM) as a new class of factor models. In QFM, both factors and loadings are allowed to be quantile-dependent objects. These extra factors could be useful for identification purposes, for instance mean-shifting factors vs. volatility/skewness/kurtosis factors, as well as for forecasting purposes in factor-augmented regressions and FAVAR setups.

Using tools in the interface of QR, Principal Component Analysis (PCA) and the theory of empirical processes, we propose a novel estimation procedure, labelled Quantile Factor Analysis (QFA), that yields consistent and asymptotically normal estimators of factors and loadings at each quantile. An important advantage of QFA is that it is able to extract simultaneously all mean-shifting and extra factors determining the factor structure of QFM, in contrast to PCA which can only extract mean-shifting factors. In addition, we propose two selection criteria to estimate consistently the number of factors at each quantile. Finally, another interesting result is that QFA estimators remain valid when the idiosyncratic error terms in AFM exhibit heavy tails and outliers, which is a case where PCA is rendered invalid.

The previous theoretical findings receive support in finite samples from a range of Monte Carlo simulations. Furthermore, it is shown in these simulations that QFA estimation performs well when we depart from some of simplifying assumptions used in the theory section for tractability (like, e.g., independence of the idiosyncratic errors). Lastly, our empirical applications to three large panel datasets of financial, macro and climate variables provide evidence that some these extra factors may be highly relevant in practice.

Any time a novel methodology is proposed, new research issues emerge for future investigation. Among the ones which have been left out of this paper (some are part of our current research agenda), four topics stand out as important:

- Factor augmented regressions and FAVAR: In relation to this topic, it would also be interesting to check the contributions of the extra factors in forecasting and monitoring (see, e.g., Stock and Watson 2002 for this type of analysis). This is an issue of high interest for applied researchers, especially with the surge of Big Data technologies. For example, one could analyze the role of the extra factors in the estimation and shock identification in FAVAR. Recent developments in quantile VAR estimation, as in White et al. (2015) provide useful tools in addressing these issues.
- Relaxing the independence assumptions: in view of the simulation results in Tables 5 and 6, we conjecture that the main theoretical results of our paper continue to hold when the error terms in QFM are allowed to have weak cross-sectional and serial dependence. Providing a formal justification for this conjecture remains high in our research agenda. As discussed in Remark 1.4, the goal here is to provide more general conditions on u_{it}

under which the sub-Gaussian type inequalities still hold.

- Dynamic QFM: Although our methodology admits factors to exhibit dependence, provided Assumption 2(i) holds, a pending issue is how to extend our results for static QFM to dynamic QFM, where the set of quantile-dependent objects include lagged factors (see Forni et al. 2000 and Stock and Watson 2011). Since our main aim in this paper has been to introduce the new class of QFM and their basic properties, for the sake of brevity, we have focused on static QFM, leaving this topic for further investigation.
- Economic interpretation of QFA factors in empirical applications: given the evidence that extra factors could be relevant in practice, another interesting issue is how to interpret them in different economic and financial contexts. Once the econometric techniques to detect and estimate extra factors in QFM have been established, attempts to provide new economic insights for these objects would help enrich the economic theory underlying this type of factor structures.

A Tables and Figures

Table 1: AFM with Cauchy-distributed Error Terms: Number of Factors

N	T	PC_{p1} of BN	IC_{p1} of BN	Eigenvalue Ratio	Rank Estimator
50	50	[0.0, 0.0, 100]	[0.1, 0.2, 99.7]	[74.6, 10.5, 14.9]	[43.2, 32.5, 24.3]
50	100	[0.0, 0.0, 100]	[0.0, 0.2, 99.8]	[75.8, 9.9, 14.3]	[37.7, 54.9, 7.4]
50	200	[0.0, 0.0, 100]	[0.0, 0.1, 99.9]	[74.0, 11.3, 14.7]	[46.3, 48.1.0, 5.6]
100	50	[0.0, 0.0, 100]	[0.0, 0.0, 100]	[76.3, 9.7, 14.0]	[39.1, 52.0, 8.9]
100	100	[0.0, 0.0, 100]	[0.0, 0.0, 100]	[75.2, 9.5, 15.3]	[8.9, 90.3, 0.9]
100	200	[0.0, 0.0, 100]	[0.0, 0.0, 100]	[74.1, 11.3, 14.6]	[7.4, 92.2, 0.4]
200	50	[0.0, 0.0, 100]	[0.0, 0.0, 100]	[75.7, 11.4, 12.9]	[41.0, 55.2, 3.8]
200	100	[0.0, 0.0, 100]	[0.0, 0.0, 100]	[74.0, 11.7, 14.3]	[7.1, 92.6, 0.3]
200	200	[0.0, 0.0, 100]	[0.0, 0.0, 100]	[72.4, 11.3, 16.3]	[0.0, 100, 0.0]

Note: The DGP considered in this Table: $X_{it} = \sum_{j=1}^{3} \lambda_{ji} f_{jt} + u_{it}$, where $f_{1t} = 0.2 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = 0.8 f_{3,t-1} + \epsilon_{3t}$, $\lambda_{ji}, \epsilon_{jt} \sim$ $i.i.d\, \mathcal{N}(0,1)$, $u_{it} \sim$ $i.i.d$ Cauchy$(0,1)$. For each estimation method, we reported [proportion of $\hat{r} < 3$, proportion of $\hat{r} = 3$, proportion of $\hat{r} > 3$] from 1000 replications.

Table 2: AFM with Cauchy-distributed Error Terms: Estimation of Factors

N	T	f_{1t},$\hat{F}_{PCA}$	f_{2t},$\hat{F}_{PCA}$	f_{3t},$\hat{F}_{PCA}$	f_{1t},$\hat{F}_{QFA}$	f_{2t},$\hat{F}_{QFA}$	f_{3t},$\hat{F}_{QFA}$
50	50	0.062	0.063	0.067	0.914	0.919	0.964
50	100	0.030	0.030	0.031	0.927	0.942	0.970
50	200	0.015	0.015	0.015	0.932	0.945	0.972
100	50	0.062	0.062	0.061	0.963	0.971	0.985
100	100	0.030	0.030	0.031	0.969	0.975	0.988
100	200	0.015	0.015	0.015	0.971	0.977	0.988
200	50	0.061	0.060	0.059	0.982	0.986	0.993
200	100	0.029	0.030	0.031	0.986	0.989	0.994
200	200	0.015	0.014	0.015	0.987	0.989	0.995

Note: The DGP considered in this Table is: $X_{it} = \sum_{j=1}^{3} \lambda_{ji} f_{jt} + u_{it}$, where $f_{1t} = 0.2 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = 0.8 f_{3,t-1} + \epsilon_{3t}$, $\lambda_{ji}, \epsilon_{jt} \sim i.i.d\, \mathcal{N}(0,1)$, $u_{it} \sim$ $i.i.d$ Cauchy$(0,1)$. For each estimation method, we report the average R^2 in the regression of (each of) the true factors on the estimated factors by PCA and QFA.

Table 3: Estimation of QFM: Independent Error Terms

		$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
N	T	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}
50	50	2.21	0.866	0.721	0.339	1.91	0.956	0.808	0.013	2.23	0.926	0.738	0.334
50	100	2.42	0.943	0.758	0.483	1.88	0.968	0.839	0.003	2.38	0.946	0.708	0.463
50	200	2.43	0.933	0.703	0.485	1.88	0.971	0.842	0.001	2.40	0.951	0.698	0.445
100	50	2.14	0.944	0.681	0.337	1.80	0.980	0.786	0.014	2.13	0.948	0.694	0.357
100	100	2.71	0.977	0.898	0.688	1.98	0.985	0.954	0.001	2.72	0.968	0.890	0.707
100	200	2.82	0.983	0.904	0.757	1.99	0.987	0.966	0.003	2.86	0.982	0.908	0.793
200	50	2.35	0.970	0.826	0.490	1.87	0.989	0.867	0.008	2.29	0.973	0.745	0.489
200	100	2.80	0.990	0.934	0.782	2.00	0.993	0.987	0.001	2.81	0.990	0.977	0.772
200	200	2.99	0.992	0.986	0.940	2.00	0.994	0.988	0.000	2.99	0.992	0.986	0.935

Note: The DGP considered in this Table is: $X_{it} = \lambda_{1i} f_{1t} + \lambda_{2i} f_{2t} + (\lambda_{3i} f_{3t}) \cdot e_{it}$, $f_{1t} = 0.8 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = |g_t|$, $\lambda_{1i}, \lambda_{2i}, \epsilon_{1t}, \epsilon_{2t}, g_t \sim i.i.d\ \mathcal{N}(0,1)$, and $\lambda_{3i} \sim i.i.d\ U[1,2]$. $e_{it} = \beta e_{i,t-1} + v_{it} + \rho \cdot \sum_{j=i-J, j\neq i}^{i+J} v_{jt}$, $v_{it} \sim i.i.d\ \mathcal{N}(0,1)$, $\beta = \rho = 0$. For each τ, the first column reports the average of $\hat{r}_{rank}$ from 1000 replications, the second to the fourth columns report the average R^2 in the regression of (each of) the true factors on the QFA factors $\hat{F}^{\hat{r}}_{QFA}$, obtained from the IQR algorithm.

Table 4: Estimation of QFM: Independent Error Terms with Heavy Tails

		$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
N	T	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}
50	50	2.81	0.911	0.727	0.585	2.38	0.954	0.827	0.031	2.95	0.925	0.711	0.617
50	100	2.79	0.934	0.782	0.621	2.03	0.963	0.885	0.005	2.79	0.933	0.783	0.658
50	200	2.82	0.942	0.811	0.680	1.91	0.966	0.855	0.000	2.76	0.943	0.790	0.648
100	50	3.20	0.962	0.851	0.737	2.67	0.977	0.907	0.076	3.07	0.942	0.828	0.682
100	100	3.06	0.972	0.897	0.840	2.21	0.983	0.939	0.018	3.06	0.974	0.931	0.801
100	200	3.00	0.974	0.944	0.867	1.99	0.983	0.958	0.000	2.98	0.974	0.943	0.860
200	50	3.24	0.971	0.839	0.753	2.82	0.984	0.903	0.106	3.31	0.970	0.858	0.773
200	100	3.10	0.985	0.937	0.897	2.31	0.991	0.975	0.018	3.09	0.987	0.949	0.883
200	200	3.02	0.989	0.977	0.932	2.07	0.992	0.985	0.005	3.02	0.988	0.978	0.933

Note: The DGP considered in this Table is: $X_{it} = \lambda_{1i} f_{1t} + \lambda_{2i} f_{2t} + (\lambda_{3i} f_{3t}) \cdot e_{it}$, $f_{1t} = 0.8 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = |g_t|$, $\lambda_{1i}, \lambda_{2i}, \epsilon_{1t}, \epsilon_{2t}, g_t \sim i.i.d\ \mathcal{N}(0,1)$, and $\lambda_{3i} \sim i.i.d\ U[1,2]$. $e_{it} = \beta e_{i,t-1} + v_{it} + \rho \cdot \sum_{j=i-J, j\neq i}^{i+J} v_{jt}$, $v_{it} \sim i.i.d$ Student(3), $\beta = \rho = 0$. For each τ, the first column reports the average of $\hat{r}_{rank}$ from 1000 replications, the second to the fourth columns report the averages of R^2 in the regression of (each of) the true factors on the QFA factors $\hat{F}^{\hat{r}}_{QFA}$, obtained from the IQR algorithm.

Table 5: Estimation of QFM: Serially Correlated Error Terms

		$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
N	T	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}
50	50	2.31	0.900	0.698	0.400	1.97	0.961	0.805	0.023	2.32	0.924	0.705	0.416
50	100	2.40	0.927	0.722	0.475	1.91	0.968	0.863	0.005	2.38	0.940	0.709	0.453
50	200	2.66	0.956	0.841	0.586	1.95	0.970	0.904	0.000	2.70	0.948	0.824	0.628
100	50	2.33	0.945	0.736	0.479	1.91	0.980	0.857	0.005	2.32	0.942	0.737	0.478
100	100	2.72	0.978	0.863	0.704	1.98	0.985	0.957	0.000	2.72	0.978	0.895	0.690
100	200	2.87	0.983	0.924	0.801	1.98	0.987	0.955	0.000	2.88	0.965	0.948	0.805
200	50	2.35	0.974	0.724	0.540	1.92	0.989	0.859	0.021	2.40	0.963	0.758	0.531
200	100	2.75	0.987	0.929	0.734	1.98	0.993	0.960	0.000	2.76	0.990	0.912	0.760
200	200	2.98	0.993	0.984	0.927	2.00	0.994	0.987	0.000	2.99	0.992	0.975	0.942

Note: The DGP considered in this Table is: $X_{it} = \lambda_{1i} f_{1t} + \lambda_{2i} f_{2t} + (\lambda_{3i} f_{3t}) \cdot e_{it}$, $f_{1t} = 0.8 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = |g_t|$, $\lambda_{1i}, \lambda_{2i}, \epsilon_{1t}, \epsilon_{2t}, g_t \sim i.i.d\ \mathcal{N}(0,1)$, and $\lambda_{3i} \sim i.i.d\ U[1,2]$. $e_{it} = \beta * e_{i,t-1} + v_{it} + \rho \cdot \sum_{j=i-J, j \neq i}^{i+J} v_{jt}$, $v_{it} \sim i.i.d\ \mathcal{N}(0,1)$, $\beta = 0.2$, $\rho = 0$. For each τ, the first column reports the average of $\hat{r}_{rank}$ from 1000 replications, the second to the fourth columns report the average R^2 in the regression of (each of) the true factors on the QFA factors $\hat{F}^{\hat{r}}_{QFA}$, obtained from the IQR algorithm.

Table 6: Estimation of QFM: Serially and Cross-Sectionally Correlated Error Terms

		$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
N	T	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}_{rank}$	f_{1t}	f_{2t}	f_{3t}
50	50	2.54	0.926	0.705	0.409	2.16	0.952	0.808	0.029	2.53	0.921	0.700	0.423
50	100	2.49	0.941	0.703	0.397	1.95	0.959	0.845	0.001	2.50	0.934	0.723	0.423
50	200	2.66	0.945	0.803	0.460	1.97	0.963	0.881	0.000	2.64	0.939	0.756	0.471
100	50	2.52	0.942	0.780	0.495	2.02	0.977	0.820	0.021	2.41	0.946	0.744	0.472
100	100	2.91	0.976	0.896	0.697	2.06	0.981	0.945	0.006	2.87	0.977	0.893	0.686
100	200	2.90	0.979	0.924	0.702	2.01	0.983	0.966	0.000	2.92	0.980	0.933	0.713
200	50	2.47	0.967	0.732	0.569	2.05	0.987	0.870	0.032	2.52	0.969	0.785	0.576
200	100	2.88	0.989	0.913	0.802	2.00	0.991	0.982	0.000	2.89	0.989	0.938	0.788
200	200	3.00	0.990	0.982	0.866	2.00	0.992	0.983	0.000	3.00	0.990	0.981	0.866

Note: The DGP considered in this Table is: $X_{it} = \lambda_{1i} f_{1t} + \lambda_{2i} f_{2t} + (\lambda_{3i} f_{3t}) \cdot e_{it}$, $f_{1t} = 0.8 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = |g_t|$, $\lambda_{1i}, \lambda_{2i}, \epsilon_{1t}, \epsilon_{2t}, g_t \sim i.i.d\ \mathcal{N}(0,1)$, and $\lambda_{3i} \sim i.i.d\ U[1,2]$. $e_{it} = \beta e_{i,t-1} + v_{it} + \rho \cdot \sum_{j=i-J, j \neq i}^{i+J} v_{jt}$, $v_{it} \sim i.i.d\ \mathcal{N}(0,1)$, $\beta = \rho = 0.2$ and $J = 3$. For each τ, the first column reports the average of $\hat{r}_{rank}$ from 1000 replications, the second to the fourth columns report the average R^2 in the regression of (each of) the true factors on the QFA factors $\hat{F}^{\hat{r}}_{QFA}$, obtained from the IQR algorithm.

Table 7: Empirical Applications: Number of Factors

	SW	Climate	MF
(N, T)	(167,221)	(338,100)	(2378,180)
No. of PCA factors	3	2	3
$\hat{r}_{\text{rank}}$ $\tau = 0.1$	2	2	3
$\hat{r}_{\text{rank}}$ $\tau = 0.25$	3	6	3
$\hat{r}_{\text{rank}}$ $\tau = 0.5$	3	6	3
$\hat{r}_{\text{rank}}$ $\tau = 0.75$	4	5	3
$\hat{r}_{\text{rank}}$ $\tau = 0.9$	2	2	3

Note: This table provides the estimated numbers of PCA factors using the eigenvalue-ratio estimator, and the estimated numbers of QFA factors at $\tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ using the rank-minimization estimator.

Table 8: Empirical Applications: Comparison of $\hat{F}_{FQR}$ with $\hat{F}_{PCA}$

	Dataset	$\hat{F}_{QFA,1}$	$\hat{F}_{QFA,2}$	$\hat{F}_{QFA,3}$	$\hat{F}_{QFA,4}$	$\hat{F}_{QFA,5}$	$\hat{F}_{QFA,6}$
$\tau = 0.1$	SW	0.745	0.850				
$\tau = 0.25$	SW	0.949	0.750	0.880			
$\tau = 0.5$	SW	0.990	0.907	0.942			
$\tau = 0.75$	SW	0.892	0.850	0.899	0.359		
$\tau = 0.9$	SW	0.135	0.919				
$\tau = 0.1$	Climate	0.581	0.010				
$\tau = 0.25$	Climate	0.955	0.955	0.000	0.544	0.031	0.000
$\tau = 0.5$	Climate	0.989	0.984	0.000	0.000	0.000	0.000
$\tau = 0.75$	Climate	0.882	0.961	0.313	0.000	0.153	
$\tau = 0.9$	Climate	0.619	0.834				
$\tau = 0.1$	MF	0.939	0.887	0.117			
$\tau = 0.25$	MF	0.980	0.983	0.038			
$\tau = 0.5$	MF	0.996	0.982	0.994			
$\tau = 0.75$	MF	0.965	0.967	0.943			
$\tau = 0.9$	MF	0.871	0.917	0.919			

Note: This table reports the R^2 of regressing each element of $\hat{F}_{QFA}$ on $\hat{F}_{PCA}$. For $\hat{F}_{QFA}$ and $\hat{F}_{PCA}$, the numbers of estimated factors is obtained from Table 7.

Table 9: Empirical Applications: Comparison of $\hat{F}_{QFA}$ with $\hat{VF}_2, \hat{VF}_3, \hat{VF}_4$.

SW	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
$\hat{VF}_2$	0.647	0.505	0.366	0.370	0.567
$\hat{VF}_3$	0.469	0.502	0.378	0.423	0.346
$\hat{VF}_4$	0.477	0.419	0.253	0.222	0.367
Climate	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
$\hat{VF}_2$	0.114	0.070	0.048	0.094	0.142
$\hat{VF}_3$	0.567	0.731	0.806	0.717	0.530
$\hat{VF}_4$	0.047	0.059	0.031	0.069	0.108
MF	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
$\hat{VF}_2$	0.178	0.076	0.112	0.151	0.213
$\hat{VF}_3$	0.814	0.862	0.888	0.884	0.857
$\hat{VF}_4$	0.198	0.085	0.047	0.055	0.107

Note: This table reports the R^2 of regressing $\hat{VF}_j$ on $\hat{F}_{QFA}$ for $j = 2, 3, 4$. For $\hat{F}_{QFA}$, the numbers of estimated factors is obtained from Table 7. $\hat{VF}_2$, $\hat{VF}_3$ and $\hat{VF}_4$ are the estimated volatility factor, skewness factor and kurtosis factor using the PCA-SQ approach and its extension to the cubes and fourth power of the residuals, respectively.

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Andersen, T., N. Fusari, and V. Todorov (2018). The pricing of tail risk and the equity premium: Evidence from international option markets. CREATES research papers, Department of Economics and Business Economics, Aarhus University.

Ando, T. and J. Bai (2018). Quantile co-movement in financial markets: A panel quantile model with unobserved heterogeneity. *Journal of the American Statistical Association* (just-accepted), 1–35.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *The Journal of Finance 61*(1), 259–299.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4), 1133–1150.

Bai, J. and S. Ng (2008a). Extremum estimation when the predictors are estimated from large panels. *Annals of Economics and Finance 9*(2), 201–222.

Bai, J. and S. Ng (2008b). *Large dimensional factor analysis.* Now Publishers Inc.

Bai, J. and S. Ng (2013). Principal components estimation and identification of static factors. *Journal of Econometrics 176*(1), 18–29.

Barigozzi, M. and M. Hallin (2016). Generalized dynamic factor models and volatilities: recovering the market volatility shocks. *The Econometrics Journal 19*(1).

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–304.

Chen, L., J. J. Dolado, and J. Gonzalo (2017). Quantile factor models. *Working paper, Universidad Carlos III de Madrid*.

Chen, M., I. Fernández-Val, and M. Weidner (2018). Nonlinear factor models for network and panel data. *Working paper, UCL.*

Chen, X., L. P. Hansen, J. Scheinkman, et al. (2009). Nonlinear principal components and long-run implications of multivariate diffusions. *The Annals of Statistics 37*(6B), 4279–4312.

Collins, M., S. Dasgupta, and R. E. Schapire (2002). A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pp. 617–624.

Connor, G. and R. A. Korajczyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics 15*(3), 373–394.

Diebold, F. X. (2012). A personal perspective on the origin(s) and development of 'big data': The phenomenon, the term, and the discipline. *Working paper, UPenn*.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics 82*(4), 540–554.

Gadea, R. and J. Gonzalo (2019). Trends in distributional characteristics: Existence of global warming. *Journal of Econometrics, forthcoming.*

Galvao, A. F. and K. Kato (2016). Smoothed quantile regression for panel data. *Journal of Econometrics 193*(1), 92–112.

Gorodnichenko, Y. and S. Ng (2017). Level and volatility factors in macroeconomic data. *Journal of Monetary Economics 91*, 52–68.

Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (2016). The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Journal of Financial Economics 119*(2), 249–283.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica*, 1327–1351.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417.

Koenker, R. (2005). *Quantile Regression*. Number 38. Cambridge University Press.

Ma, S., O. Linton, J. Gao, et al. (2017). Estimation and inference in semiparametric quantile factor models. Technical report, Monash University, Department of Econometrics and Business Statistics.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*(2), 244–258.

Pelger, M. and R. Xiong (2018). Interpretable proximate factors for large dimensions. *arXiv preprint arXiv:1805.03373*.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Renault, E., T. Van Der Heijden, and B. Werker (2017). Arbitrage pricing theory for idiosyncratic variance factors. *Working paper, Brown University*.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3), 341–360.

Stock, J. H. and M. Watson (2016). Factor models and structural vector autoregressions in macroeconomics. *Handbook of Marcroeconomics 2A*, 415 – 526.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Stock, J. H. and M. W. Watson (2011). Dynamic factor models. *Oxford Handbook of Economic Forecasting 1*, 35–59.

Su, L. and X. Wang (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics 198*(1), 84–101.

van de Geer, S. A. (2002). On Hoeffdings inequality for dependent random variables. In *Empirical process techniques for dependent data*, pp. 161–169. Springer.

Van der Vaart, A. and J. Wellner (1996). *Weak convergence and empirical processes*. Springer, New York.

White, H., T.-H. Kim, and S. Manganelli (2015). VAR for VaR: Measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics 187*(1), 169–188.

Online Appendix to "Quantile Factor Models"

Liang Chen[1], Juan J. Dolado[2], and Jesús Gonzalo[3]

[1] *School of Economics, Shanghai University of Finance and Economics, chen.liang@mail.shufe.edu.cn*
[2] *Department of Economics, Universidad Carlos III de Madrid, dolado@eco.uc3m.es*
[3] *Department of Economics, Universidad Carlos III de Madrid, jgonzalo@est-eco.uc3m.es*

April 20, 2019

A Proofs of the Main Results

Definitions and Notations: Throughout the appendix, $K_1, K_2, \ldots, K_{18}$ denote some positive constants that do not depend on N, T. For any random variable Y, define the Orlicz norm $\|Y\|_\psi$ as:

$$\|Y\|_\psi = \inf\left\{C > 0 : \mathbb{E}\psi\left(|Y|/C\right) \le 1\right\},$$

where ψ is a nondecreasing, convex function with $\psi(0) = 0$. In particular, when $\psi(x) = e^{x^2} - 1$, the norm is written as $\|Y\|_{\psi_2}$. We use $\|\cdot\|$, $\|\cdot\|_S$ and $\|\cdot\|_{\max}$ to denote the Frobenius norm, the spectral norm, and the max norm for matrices, respectively. Notice that, when considering vectors, $\|\cdot\|$ is the Euclidean norm. For a matrix A with real eigenvalues, let $\rho_j(A)$ denote the j-th largest eigenvalue. Following Van der Vaart and Wellner (1996), the symbol $\lesssim$ means "left side bounded by a positive constant times the right side" (the symbol $\gtrsim$ is defined similarly), and $D(\cdot, g, \mathcal{G})$ and $C(\cdot, g, \mathcal{G})$ denote the packing and covering numbers of space $\mathcal{G}$ endowed with metric g.

A.1 Proof of Theorem 1

Define

$$\tilde{\mathbb{M}}_{NT}(\theta) = \mathbb{M}_{NT}(\theta) - \bar{\mathbb{M}}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\rho_\tau(X_{it} - \lambda_i' f_t) - \mathbb{E}[\rho_\tau(X_{it} - \lambda_i' f_t)]\right\},$$

$$\mathbb{W}_{NT}(\theta) = \tilde{\mathbb{M}}_{NT}(\theta) - \tilde{\mathbb{M}}_{NT}(\theta_0).$$

Lemma 1. *Under Assumption 1, $d(\hat{\theta}, \theta_0) = o_P(1)$ as $N, T \to \infty$.*

Proof. First, expanding $\mathbb{E}[\rho_\tau(X_{it}-c)]$ around $c_{0,it}=\lambda_{0i}f_{0t}$, we have

$$\mathbb{E}[\rho_\tau(X_{it}-c)]=\mathbb{E}[\rho_\tau(X_{it}-\lambda_{0i}f_{0t})]+0.5\cdot\mathsf{f}_{it}(c^*)\cdot(c-\lambda_{0i}f_{0t})^2,$$

where c^* is between c and $c_{0,it}$. It then follows from Assumption 1(ii) that for all $\lambda_i\in\mathcal{A}$ and $f_t\in\mathcal{F}$,

$$\mathbb{E}[\rho_\tau(X_{it}-\lambda_i'f_t)]-\mathbb{E}[\rho_\tau(X_{it}-\lambda_{0i}f_{0t})]\gtrsim(\lambda_i'f_t-\lambda_{0i}'f_{0t})^2$$

since $|\lambda_i'f_t|$ and $|\lambda_{0i}'f_{0t}|$ are both bounded. Therefore, for any $\theta\in\Theta^M$:

$$\bar{\mathbb{M}}_{NT}(\theta)-\bar{\mathbb{M}}_{NT}(\theta_0)\gtrsim d^2(\theta,\theta_0). \tag{A.1}$$

Second, by the definition of $\hat\theta$, $\mathbb{M}_{NT}(\hat\theta)-\mathbb{M}_{NT}(\theta_0)\le 0$, or equivalently $\mathbb{W}_{NT}(\hat\theta)+\bar{\mathbb{M}}_{NT}(\hat\theta)-\bar{\mathbb{M}}_{NT}(\theta_0)\le 0$. It then follows from (A.1) that

$$0\le d^2(\hat\theta,\theta_0)\lesssim\bar{\mathbb{M}}_{NT}(\hat\theta)-\bar{\mathbb{M}}_{NT}(\theta_0)\le\sup_{\theta\in\Theta^M}|\mathbb{W}_{NT}(\theta)|.$$

Thus, it remains to show that

$$\sup_{\theta\in\Theta^M}|\mathbb{W}_{NT}(\theta)|=o_P(1). \tag{A.2}$$

Choose K_1 large enough such that $\|\lambda_{0i}\|,\|f_{0t}\|,\|\lambda_i\|,\|f_t\|\le K_1$ for all i,t for any $\theta\in\Theta^M$. Let $B_r(K_1)$ be a Euclidean ball in $\mathbb{R}^r$ with radius K_1. For any $\epsilon>0$, let $\lambda_{(1)},\ldots,\lambda_{(J)}$ be a maximal set of points in $B_r(K_1)$ such that $\|\lambda_{(j)}-\lambda_{(h)}\|>\epsilon/K_1$ for any $j\ne h$. Similarly, let $f_{(1)},\ldots,f_{(J)}$ be a maximal set of points in $B_r(K_1)$ such that $\|f_{(j)}-f_{(h)}\|>\epsilon/K_1$ for any $j\ne h$. It is well known that J, the packing number of $B_r(K_1)$, is equal to $K_2(K_1/\epsilon)^r$.

For any $\theta\in\Theta^M$, define $\theta^*=(\lambda_1^{*'},\ldots,\lambda_N^{*'},f_1^{*'},\ldots,f_T^{*'})'$, where $\lambda_i^*=\{\lambda_{(j)}:j\le J,\|\lambda_{(j)}-\lambda_i\|\le\epsilon/K_1\}$ and $f_t^*=\{f_{(j)}:j\le J,\|f_{(j)}-f_t\|\le\epsilon/K_1\}$. Thus, we can write

$$\mathbb{W}_{NT}(\theta)=\mathbb{W}_{NT}(\theta^*)+\mathbb{W}_{NT}(\theta)-\mathbb{W}_{NT}(\theta^*).$$

Note that $|\rho_\tau(X_{it}-\lambda_i'f_t)-\rho_\tau(X_{it}-\lambda_i^{*'}f_t^*)|\le 2|\lambda_i'f_t-\lambda_i^{*'}f_t^*|\le 2\|\lambda_i\|\|f_t-f_t^*\|+2\|f_t^*\|\|\lambda_i-\lambda_i^*\|\le 4\epsilon$. Thus,

$$\sup_{\theta\in\Theta^M}|\mathbb{W}_{NT}(\theta)-\mathbb{W}_{NT}(\theta^*)|\lesssim\epsilon. \tag{A.3}$$

Also, note that $|\rho_\tau(X_{it}-\lambda_{0i}'f_{0t})-\rho_\tau(X_{it}-\lambda_i^{*'}f_t^*)|\le 2|\lambda_{0i}'f_{0t}-\lambda_i^{*'}f_t^*|$. Then, by Hoeffding's inequality

$$P[|\sqrt{NT}\mathbb{W}_{NT}(\theta^*)|>c]\le 2e^{-\frac{2c^2}{4\cdot d^2(\theta^*,\theta_0)}},$$

and by Lemma 2.2.1 of Van der Vaart and Wellner (1996) it follows that $\|\mathbb{W}_{NT}(\theta^*)\|_{\psi_2}\lesssim d(\theta^*,\theta_0)/\sqrt{NT}$. Since θ^* can take at most $J^{N+T}\lesssim(K_1/\epsilon)^{r(N+T)}$ different values, and $d(\theta^*,\theta_0)\le 2K_1$, it follows from Lemma 2.2.2 of Van der Vaart and Wellner (1996) that

$$\mathbb{E}\left[\sup_{\theta\in\Theta^M}|\mathbb{W}_{NT}(\theta^*)|\right]\lesssim\sqrt{\log(K_1/\epsilon)}\sqrt{r(N+T)}/\sqrt{NT}\lesssim\sqrt{\log(K_1/\epsilon)}/L_{NT}. \tag{A.4}$$

Finally, from (A.3) and (A.4)

$$\mathbb{E}\left[\sup_{\theta\in\Theta^M}|\mathbb{W}_{NT}(\theta)|\right]\leq\mathbb{E}\left[\sup_{\theta\in\Theta^M}|\mathbb{W}_{NT}(\theta^*)|\right]+\mathbb{E}\left[\sup_{\theta\in\Theta^M}|\mathbb{W}_{NT}(\theta)-\mathbb{W}_{NT}(\theta^*)|\right]\lesssim\sqrt{\log(K_1/\epsilon)}/L_{NT}+\epsilon.$$

Then (A.2) is satisfied since ϵ is arbitrary. This concludes the proof. $\square$

Define $\Theta^M(\delta)=\{\theta\in\Theta^M : d(\theta,\theta_0)\leq\delta\}$.

Lemma 2. *Under Assumption 1, for sufficiently small $\delta>0$,*

$$\|\Lambda-\Lambda_0\|/\sqrt{N}+\|F-F_0\|/\sqrt{T}\leq K_3\delta$$

for any $\theta\in\Theta^M(\delta)$.

Proof. Since $F'F/T=F_0'F_0/T=\mathbb{I}_r$, and $\|\Lambda_0\|/\sqrt{N}\leq K_4$ by Assumption 1(i),

$$\begin{aligned}
&\|\Lambda-\Lambda_0\|/\sqrt{N}=\|(\Lambda-\Lambda_0)F'\|/\sqrt{NT}=\|\Lambda F'-\Lambda_0F_0'+\Lambda_0F_0'-\Lambda_0F'\|/\sqrt{NT}\\
\leq\ &\|\Lambda F'-\Lambda_0F_0'\|/\sqrt{NT}+\|\Lambda_0\|/\sqrt{N}\cdot\|F-F_0\|/\sqrt{T}\\
\leq\ &d(\theta,\theta_0)+K_4\|F-F_0\|/\sqrt{T}.
\end{aligned}$$

Thus, for $\theta\in\Theta^M(\delta)$,

$$\|\Lambda-\Lambda_0\|/\sqrt{N}+\|F-F_0\|/\sqrt{T}\leq\delta+(1+K_4)\|F-F_0\|/\sqrt{T}. \tag{A.5}$$

Next,

$$\begin{aligned}
\|F-F_0\|/\sqrt{T}&\leq\|F_0-F(F'F_0/T)\|/\sqrt{T}+\|F(F'F_0/T)-F\|/\sqrt{T}\\
&=\|M_FF_0\|/\sqrt{T}+\|(F'F_0/T)-\mathbb{I}_r\|,
\end{aligned}\tag{A.6}$$

where $P_A=A(A'A)^{-1}A'$ and $M_A=\mathbb{I}-P_A$.

Third,

$$\begin{aligned}
\frac{1}{\sqrt{NT}}\|(\Lambda F'-\Lambda_0F_0')M_F\|&\leq\sqrt{\operatorname{rank}[(\Lambda F'-\Lambda_0F_0')M_F]}\cdot\|M_F\|_S\cdot\|\Lambda F'-\Lambda_0F_0'\|_S/\sqrt{NT}\\
&\lesssim\|\Lambda F'-\Lambda_0F_0'\|/\sqrt{NT},
\end{aligned}\tag{A.7}$$

and since

$$\begin{aligned}
\|(\Lambda F'-\Lambda_0F_0')M_F\|/\sqrt{NT}=\|\Lambda_0F_0'M_F\|/\sqrt{NT}&=\sqrt{\operatorname{Tr}\left[(\Lambda_0'\Lambda_0/N)\cdot(F_0'M_FF_0/T)\right]}\\
&\geq\sqrt{\sigma_{Nr}}\sqrt{\operatorname{Tr}\left(F_0'M_FF_0/T\right)}=\sqrt{\sigma_{Nr}}\|M_FF_0\|/\sqrt{T},
\end{aligned}\tag{A.8}$$

it follows from (A.7) and (A.8) that

$$\|M_F F_0\|/\sqrt{T} \lesssim \sqrt{\frac{1}{\sigma_{Nr}}} d(\theta, \theta_0). \tag{A.9}$$

Similarly, it can be shown that

$$\|M_{F_0} F\|/\sqrt{T} \lesssim \sqrt{\frac{1}{\rho_{\min}(\Lambda'\Lambda/N)}} d(\theta, \theta_0), \tag{A.10}$$

where $\rho_{\min}$ denotes the minimum eigenvalue.

Fourth,

$$\frac{1}{\sqrt{NT}}\|(\Lambda F' - \Lambda_0 F_0')P_F\| \leq \frac{1}{\sqrt{NT}}\|\Lambda F' - \Lambda_0 F_0'\| \cdot \|P_F\| = \sqrt{r} d(\theta, \theta_0),$$

so

$$\frac{1}{\sqrt{NT}}\|(\Lambda F' - \Lambda_0 F_0')P_F\| = \frac{1}{\sqrt{N}}\|\Lambda - \Lambda_0(F_0'F/T)\| \leq \sqrt{r} d(\theta, \theta_0). \tag{A.11}$$

Similarly, we can show that

$$\frac{1}{\sqrt{N}}\|\Lambda_0 - \Lambda(F'F_0/T)\| \leq \sqrt{r} d(\theta, \theta_0). \tag{A.12}$$

Next, define $R_T = F'F_0/T$. Note that $FR_T = FF'F_0/T = P_F F_0$. Then,

$$\mathbb{I}_r = F_0'F_0/T = R_T'(F'F/T)R_T + (F_0'F_0/T - R_T'(F'F/T)R_T) = R_T'R_T + F_0'M_F F_0/T, \tag{A.13}$$

and

$$\begin{aligned}\Lambda_0'\Lambda_0/N &= R_T'(\Lambda'\Lambda/N)R_T + (\Lambda_0'\Lambda_0/N - R_T'(\Lambda'\Lambda/N)R_T) \\ &= R_T'(\Lambda'\Lambda/N)R_T + \Lambda_0'(\Lambda_0 - \Lambda R_T)/N + (\Lambda_0 - \Lambda R_T)'\Lambda R_T/N. \end{aligned} \tag{A.14}$$

Similarly,

$$\mathbb{I}_r = R_T R_T' + F'M_{F_0}F/T. \tag{A.15}$$

From (A.14),

$$\begin{aligned} &\Lambda_0'\Lambda_0/N = R_T'(\Lambda'\Lambda/N)(R_T')^{-1}R_T'R_T + \Lambda_0'(\Lambda_0 - \Lambda R_T)/N + (\Lambda_0 - \Lambda R_T)'\Lambda R_T/N \\ = \; &R_T'(\Lambda'\Lambda/N)(R_T')^{-1} + R_T'(\Lambda'\Lambda/N)(R_T')^{-1}(R_T'R_T - \mathbb{I}_r) + \Lambda_0'(\Lambda_0 - \Lambda R_T)/N + (\Lambda_0 - \Lambda R_T)'\Lambda R_T/N, \end{aligned}$$

and it follows from the above equation and (A.13) that

$$(\Lambda_0'\Lambda_0/N + D_{NT})\, R_T' = R_T'(\Lambda'\Lambda/N), \tag{A.16}$$

where

$$D_{NT} = R_T'(\Lambda'\Lambda/N)(R_T')^{-1}F_0'M_F F_0/T - \Lambda_0'(\Lambda_0 - \Lambda R_T)/N - (\Lambda_0 - \Lambda R_T)'\Lambda R_T/N.$$

From (A.9) and (A.12) we have that $\|D_{NT}\| \lesssim d(\theta, \theta_0)$. By matrix perturbation theory, when $d(\theta, \theta_0)$ is sufficiently small, we have

$$|\rho_{\min}[\Lambda'\Lambda/N] - \rho_{\min}[\Lambda_0'\Lambda_0/N]| \lesssim d(\theta,\theta_0), \tag{A.17}$$

$$\|R_T'V_T - \mathbb{I}_r\| \lesssim d(\theta,\theta_0), \tag{A.18}$$

where $V_T = \text{diag}\left((R_{T,1}R_{T,1}')^{-1/2}, \ldots, (R_{T,r}R_{T,r}')^{-1/2}\right)$, and $R_{T,j}'$ is the jth column of R_T'.

(A.10) and (A.17) imply that

$$\|M_{F_0}F\|/\sqrt{T} \lesssim d(\theta,\theta_0). \tag{A.19}$$

Note that by the triangular inequality, it holds that

$$\|R_T' - \mathbb{I}_r\| \leq \|R_T'V_T - \mathbb{I}_r\| + \|R_T'V_T - R_T'\| \leq \|R_T'V_T - \mathbb{I}_r\| + \|R_T\| \cdot \|V_T - \mathbb{I}_r\|. \tag{A.20}$$

From (A.15) and (A.19),

$$\|V_T - \mathbb{I}_r\| \leq \|R_TR_T' - \mathbb{I}_r\| \lesssim d^2(\theta,\theta_0). \tag{A.21}$$

It then follows from (A.18) (A.20) and (A.21) that for small enough $d(\theta,\theta_0)$,

$$\|R_T - \mathbb{I}_r\| \lesssim d(\theta,\theta_0). \tag{A.22}$$

Finally, it is obtained from (A.6) (A.9) and (A.22) that for sufficiently small $d(\theta,\theta_0)$

$$\|F - F_0\|/\sqrt{T} \lesssim d(\theta,\theta_0). \tag{A.23}$$

Then the desired result follows from (A.5) and (A.23). □

Lemma 3. *Under Assumption 1, for sufficiently small* δ*,*

$$\mathbb{E}\left[\sup_{\theta\in\Theta^M(\delta)} |\mathbb{W}_{NT}(\theta)|\right] \lesssim \frac{\delta}{L_{NT}}.$$

Proof. In the proof of Lemma 1 we have shown that

$$\left\|\sqrt{NT}|\mathbb{W}_{NT}(\theta_a) - \mathbb{W}_{NT}(\theta_b)|\right\|_{\psi_2} \lesssim d(\theta_a,\theta_b), \tag{A.24}$$

and therefore

$$\left\|\sqrt{NT}\mathbb{W}_{NT}(\theta)\right\|_{\psi_2} \lesssim d(\theta,\theta_0).$$

Construct nested sets $\Theta_1^M(\delta) \subset \Theta_2^M(\delta) \ldots \subset \Theta^M(\delta)$ such that each $\Theta_j^M(\delta)$ is a maximal set of points such that $d(\theta_a,\theta_b) > \delta/2^j$ for every $\theta_a \neq \theta_b$ in $\Theta_j^M(\delta)$. In particular, let $\Theta_0^M(\delta) = \{\theta_0\}$.

For each point θ in $\Theta_{j+1}^M(\delta)$, let θ_* be a point in $\Theta_j^M(\delta)$ such that $d(\theta,\theta_*) \leq \delta/2^j$. It then follows by the triangle inequality that

$$\max_{\Theta_{j+1}^M(\delta)} |\mathbb{W}_{NT}(\theta)| \leq \max_{\Theta_{j+1}^M(\delta)} |\mathbb{W}_{NT}(\theta_*)| + \max_{\Theta_{j+1}^M(\delta)} |\mathbb{W}_{NT}(\theta) - \mathbb{W}_{NT}(\theta_*)|. \tag{A.25}$$

Note that

$$\max_{\Theta^M_{j+1}(\delta)} |\mathbb{W}_{NT}(\theta_*)| \leq \max_{\Theta^M_j(\delta)} |\mathbb{W}_{NT}(\theta)|,$$

and taking $\|\|_{\psi_2}$ norm on both sides of (A.25) gives

$$\left\|\max_{\Theta^M_{j+1}(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right\|_{\psi_2} \leq \left\|\max_{\Theta^M_j(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right\|_{\psi_2} + \left\|\max_{\Theta^M_{j+1}(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta) - \mathbb{W}_{NT}(\theta_*)|\right\|_{\psi_2}.$$

Let $m_j = \#\Theta^M_j$, the number of points in Θ^M_j. The second term on the RHS of the last inequality, according to Lemma 2.2.2 of Van der Vaart and Wellner (1996), is bounded by

$$K_5\sqrt{\log(1+m_{j+1})} \cdot \max_{\Theta^M_{j+1}(\delta)} \left\|\sqrt{NT}\,|\mathbb{W}_{NT}(\theta) - \mathbb{W}_{NT}(\theta_*)|\right\|_{\psi_2},$$

which according to (A.24) is bounded by $\delta/2^j \cdot \sqrt{\log(1+m_{j+1})}$ multiplied by a positive constant. Thus we have

$$\left\|\max_{\Theta^M_{j+1}(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right\|_{\psi_2} \leq \left\|\max_{\Theta^M_j(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right\|_{\psi_2} + K_6\delta/2^j \cdot \sqrt{\log(1+m_{j+1})},$$

which implies that for $J > 1$,

$$\left\|\max_{\Theta^M_{J+1}(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right\|_{\psi_2} \leq K_6 \sum_{j=1}^{J} \delta/2^j \cdot \sqrt{\log(1+m_{j+1})}.$$

Let $J \to \infty$, the above inequality gives

$$\left\|\sup_{\Theta^M(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right\|_{\psi_2} \lesssim \sum_{j=1}^{\infty} \delta/2^j \cdot \sqrt{\log(m_{j+1})}. \tag{A.26}$$

Note that $m_{j+1} \leq D(\delta/2^{j+1}, d, \Theta^M(\delta))$, which is the packing number of $\Theta^M(\delta)$. Further note that

$$\delta/2^j \cdot \sqrt{\log\left(D(\delta/2^{j+1}, d, \Theta^M(\delta))\right)} \lesssim \int_{\delta/2^{j+2}}^{\delta/2^{j+1}} \sqrt{\log\left(D(\epsilon, d, \Theta^M(\delta))\right)}d\epsilon.$$

It follows from the above inequality that the RHS of (A.26) is bounded by $\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^M(\delta))}$, and we finally have

$$\mathbb{E}\left[\sup_{\Theta^M(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right] \lesssim \left\|\sup_{\Theta^M(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right\|_{\psi_2} \lesssim \int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^M(\delta))}d\epsilon.$$

Then, it remains to show that

$$\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^M(\delta))}d\epsilon = O(\sqrt{N+T}\delta). \tag{A.27}$$

To prove (A.27), first note that for any $\theta \in \Theta^M$,

$$d(\theta,\theta_0) = \frac{1}{\sqrt{NT}}\|\Lambda F' - \Lambda_0' F_0'\| = \frac{1}{\sqrt{NT}}\|\Lambda F' - \Lambda_0 F' + \Lambda_0 F' - \Lambda_0' F_0'\|$$
$$\leq \frac{1}{\sqrt{N}}\|\Lambda - \Lambda_0\| + \frac{\|\Lambda_0\|}{\sqrt{N}} \cdot \frac{\|F - F_0\|}{\sqrt{T}} \leq K_7\left(\frac{\|\Lambda - \Lambda_0\|}{\sqrt{N}} + \frac{\|F - F_0\|}{\sqrt{T}}\right),$$

where $K_7 \geq 1$. Now define

$$d^*(\theta,\theta_0) = 2K_7\sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^r (\lambda_{ij} - \lambda_{0,ij})^2}{N} + \frac{\sum_{t=1}^T \sum_{j=1}^r (f_{tj} - f_{0,tj})^2}{T}}.$$

Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \leq 2\sqrt{a+b}$, $d(\theta,\theta_0) \leq d^*(\theta,\theta_0)$, and by Lemma 2, $\theta \in \Theta^M(\delta)$ implies $d^*(\theta,\theta_0) \leq K_8 d(\theta,\theta_0)$ with $K_8 = 2K_7 * K_3$. Thus $\Theta^M(\delta) \subset \Theta^{M*}(\delta)$, where $\Theta^{M*}(\delta) = \{\theta \in \Theta^M : d^*(\theta,\theta_0) \leq K_8\delta\}$. It then follows that

$$D(\epsilon, d, \Theta^M(\delta)) \leq D(\epsilon, d^*, \Theta^{M*}(\delta)) \leq C(\epsilon/2, d^*, \Theta^{M*}(\delta)). \tag{A.28}$$

Next, we calculate an upper bound for $C(\epsilon/2, d^*, \Theta^{M*}(\delta))$. Let $\eta = \epsilon/2$, and $\theta_1^*, \ldots, \theta_J^*$ be a largest set in $\Theta^{M*}(\delta)$ such that $d^*(\theta_j^*, \theta_l^*) > \eta$ for any $j \neq l$. Define $B(\theta, c) = \{\gamma \in \Theta^M : d^*(\gamma,\theta) \leq c\}$. Then, the balls $B(\theta_1^*, \eta), \ldots, B(\theta_J^*, \eta)$ cover $\Theta^{M*}(\delta)$, and thus $C(\epsilon/2, d^*, \Theta^{M*}(\delta)) \leq J$. Moreover, the balls $B(\theta_1^*, \eta/4), \ldots, B(\theta_J^*, \eta/4)$ are disjoint and

$$\cup_{j=1}^J B(\theta_j^*, \eta/4) \subset \Theta^{M*}(\delta + \eta/4).$$

Note that the volume of a ball defined by the metric d^* with radius c is the volume of an ellipsoid, which is equal to $h_M \cdot c^M$, where h_M is a constant that depends on N, T and r, but not on c. Therefore,

$$J \cdot h_M \cdot (\eta/4)^M \leq h_M \cdot (K_8\delta + \eta/4)^M,$$

which implies

$$J \leq \left(\frac{4K_8\delta + \eta}{\eta}\right)^M = \left(\frac{8K_8\delta + \epsilon}{\epsilon}\right)^M \leq \left(\frac{K_9\delta}{\epsilon}\right)^M \tag{A.29}$$

for $\epsilon \leq \delta$, where $K_9 = 8K_8 + 1$. Then from (A.28) and (A.29)

$$\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^M(\delta))}d\epsilon \quad \leq \quad \int_0^\delta \sqrt{\log C(\epsilon, d^*, \Theta^{M*}(\delta))}d\epsilon \quad \leq \quad \sqrt{(N+T)r}\int_0^\delta \sqrt{\log(K_9\delta/\epsilon)}d\epsilon.$$

It is easy to show that $\int_0^\delta \sqrt{\log(K_9\delta/\epsilon)}d\epsilon = O(\delta)$ and thus (A.27) is satisfied. This concludes the proof of Lemma 3. □

Proof of Theorem 1:

Proof. The parameter space Θ^M can be partitioned into shells $S_j = \{\theta \in \Theta^M : 2^{j-1} < L_{NT} \cdot d(\theta,\theta_0) \leq 2^j\}$. If $L_{NT} \cdot d(\hat{\theta},\theta_0)$ is larger than 2^V for a given integer V, then $\hat{\theta}$ is in one of the shells S_j with $j \geq V$. In that case the infimum of the map $\theta \mapsto \mathbb{M}_{NT}(\theta) - \mathbb{M}_{NT}(\theta_0)$ over this shell is nonpositive by the definition of $\hat{\theta}$. Conclude that, for every $\eta > 0$,

$$P\left[L_{NT} \cdot d(\hat{\theta}, \theta_0) > 2^V\right] \leq \sum_{j \geq V, 2^j \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} (\mathbb{M}_{NT}(\theta) - \mathbb{M}_{NT}(\theta_0)) \leq 0\right] + P[d(\hat{\theta}, \theta_0) \geq \eta].$$

For arbitrarily small $\eta > 0$, the second probability on the RHS of the above equation converges to 0 as $N, T \to \infty$ by Lemma 1.

Next, note that by (A.1), for each θ in S_j,

$$-[\bar{\mathbb{M}}_{NT}(\theta) - \bar{\mathbb{M}}_{NT}(\theta_0)] \lesssim -d^2_{NT}(\theta, \theta_0) \leq -\frac{2^{2j-2}}{L^2_{NT}}.$$

Thus, $\inf_{\theta \in S_j} (\mathbb{M}_{NT}(\theta) - \mathbb{M}_{NT}(\theta_0)) \leq 0$ implies that

$$\inf_{\theta \in S_j} \mathbb{W}_{NT}(\theta) \leq -\frac{2^{2j-2}}{L^2_{NT}},$$

and therefore

$$\sum_{j \geq V, 2^j \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} (\mathbb{M}_{NT}(\theta) - \mathbb{M}_{NT}(\theta_0)) \leq 0\right] \leq \sum_{j \geq V, 2^j \leq \eta L_{NT}} P\left[\sup_{\theta \in S_j} |\mathbb{W}_{NT}(\theta)| \geq \frac{2^{2j-2}}{L^2_{NT}}\right].$$

By Lemma 3 and Markov's inequality, we have

$$P\left[\sup_{\theta \in S_j} |\mathbb{W}_{NT}(\theta)| \geq \frac{2^{2j-2}}{L^2_{NT}}\right] \leq \frac{L^2_{NT}}{2^{2j-2}} \cdot \mathbb{E}\left[\sup_{\theta \in S_j} |\mathbb{W}_{NT}(\theta)|\right] \lesssim \frac{L^2_{NT}}{2^{2j}} \cdot \frac{2^j}{L^2_{NT}} = 2^{-j},$$

which implies that

$$\sum_{j \geq V, 2^j \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} (\mathbb{M}_{NT}(\theta) - \mathbb{M}_{NT}(\theta_0)) \leq 0\right] \lesssim \sum_{j \geq V} 2^{-j}.$$

The RHS of the previous expression convergences to 0 as $V \to \infty$, implying that $L_{NT} \cdot d(\hat{\theta}, \theta_0) = O_P(1)$, or $d(\hat{\theta}, \theta_0) = O_P(1/L_{NT})$. The desired result then follows from Lemma 2. $\square$

A.2 Proof of Theorem 2

For sufficiently small δ, define $\Theta^k(\delta) = \{\theta^k \in \Theta^k : d(\theta^k, \theta_0) \leq \delta\}$. Let $F^{k,r}$ denote the first r columns of F^k, and let $F^{k,-r}$ denote the remaining $k - r$ columns of F^k. $\Lambda^{k,r}$ and $\Lambda^{k,-r}$ are defined similarly.

Lemma 4. *Suppose that Assumption 1 holds and $r < k < \infty$. Then for any $\theta^k \in \Theta^k(\delta)$ and sufficiently* *small* δ,

$$\|F^{k,r}-F_0\|/\sqrt{T}\lesssim\delta,\quad \|\Lambda^{k,r}-\Lambda_0\|/\sqrt{N}\lesssim\delta,\quad \|\Lambda^{k,-r}\|/\sqrt{N}\lesssim\delta.$$

Proof. First, similar to (A.9) and (A.10), it can be shown that for any $\theta^k\in\Theta^k(\delta)$,

$$\|M_{F^k}F_0\|/\sqrt{T}=\|F_0-F^k(F^{k'}F_0/T)\|/\sqrt{T}\leq\sqrt{\frac{2k}{\sigma_{Nr}}}\cdot d(\theta^k,\theta_0), \tag{A.30}$$

$$\sqrt{\mathrm{Tr}\left[(\Lambda^{k'}\Lambda^k/N)\cdot(F^{k'}M_{F_0}F^k/T)\right]}\leq\sqrt{2k}\cdot d(\theta^k,\theta_0). \tag{A.31}$$

Similar to (A.11) and (A.12) we can show that

$$\frac{1}{\sqrt{N}}\|\Lambda^k-\Lambda_0(F_0'F^k/T)\|\leq\sqrt{k}d(\theta,\theta_0),\quad \frac{1}{\sqrt{N}}\|\Lambda_0-\Lambda^k(F^{k'}F_0/T)\|\leq\sqrt{r}d(\theta,\theta_0). \tag{A.32}$$

With a little abuse of notation, define $R_T=F^{k'}F_0/T$. From the above inequalities, we have

$$\|(\Lambda^k)'\Lambda^k/N-R_T(\Lambda_0'\Lambda_0/N)R_T'\|\leq\left(\frac{\|\Lambda_0R_T'\|}{\sqrt{N}}+\frac{\|\Lambda^k\|}{\sqrt{N}}\right)\cdot\frac{\|\Lambda^k-\Lambda_0R_T'\|}{\sqrt{N}}\lesssim d(\theta,\theta_0). \tag{A.33}$$

Note that the matrix $R_T(\Lambda_0'\Lambda_0/N)R_T'$ has rank less or equal to r. Thus, for small enough δ, according to the matrix perturbation theory,

$$\rho_{r+j}\left((\Lambda^k)'\Lambda^k/N\right)-\rho_{r+j}\left(R_T(\Lambda_0'\Lambda_0/N)R_T'\right)=\sigma^k_{N,r+j}\lesssim d(\theta^k,\theta_0)\text{ for }1\leq j\leq k-r, \tag{A.34}$$

where $(\Lambda^k)'\Lambda^k/N=\mathrm{diag}\left(\sigma^k_{N,1},\ldots,\sigma^k_{N,k}\right)$.

Define $R_T^r=F^{k,r'}F_0/T$ and $R_T^{-r}=(F^{k,-r})'F_0/T$, then $R_T=(R_T^{r'},R_T^{-r'})'$. It then follows from (A.33) and (A.34) that $\|R_T^{-r}(\Lambda_0'\Lambda_0/N)R_T^{-r'}\|\lesssim d(\theta,\theta_0)$, which in turn implies that

$$\|R_T^{-r}\|^2\lesssim\|R_T^{-r}\|^2_{\max}\lesssim d(\theta^k,\theta_0). \tag{A.35}$$

Next, we can write

$$\mathbb{I}_r=F_0'F_0/T=R_T'R_T+F_0'F_0/T-R_T'(F^{k'}F^k/T)R_T=R_T^{r'}R_T^r+R_T^{-r'}R_T^{-r}+\frac{F_0'}{\sqrt{T}}\cdot\frac{F_0-F^kR_T}{\sqrt{T}}.$$

$$\begin{aligned}\Lambda_0'\Lambda_0/N=R_T'(\Lambda^{k'}\Lambda^k/N)R_T+\Lambda_0'\Lambda_0/N-R_T'(\Lambda^{k'}\Lambda^k/N)R_T=R_T^{r'}\left(\mathrm{diag}(\sigma^k_{N,1},\ldots,\sigma^k_{N,r})\right)R_T^r\\+R_T^{-r'}\left(\mathrm{diag}(\sigma^k_{N,r+1},\ldots,\sigma^k_{N,k})\right)R_T^{-r}+\Lambda_0'(\Lambda_0-\Lambda^kR_T)/N+(\Lambda_0-\Lambda^kR_T)'\Lambda R_T/N.\end{aligned}$$

Then, similar to (A.16), we can write

$$(\Lambda_0'\Lambda_0/N+D_{NT})R_T^{r'}=R_T^{r'}\left(\mathrm{diag}(\sigma^k_{N,1},\ldots,\sigma^k_{N,r})\right), \tag{A.36}$$

where

$$D_{NT} = R_T^{-r'} R_T^{-r} + \frac{F_0'}{\sqrt{T}} \cdot \frac{F_0 - F^k R_T}{\sqrt{T}} - R_T^{-r'} \left(\mathrm{diag}(\sigma_{N,r+1}^k, \ldots, \sigma_{N,k}^k)\right) R_T^{-r} \\ - \Lambda_0'(\Lambda_0 - \Lambda^k R_T)/N - (\Lambda_0 - \Lambda^k R_T)'\Lambda R_T/N,$$

and it follows from (A.30) to (A.35) that $\|D_{NT}\| \lesssim d(\theta^k, \theta_0)$. Therefore, similar to the proof of Lemma 2, we have

$$\|R_T^r - \mathbb{I}_r\| \lesssim d(\theta^k, \theta_0) \quad \text{and} \quad \|\sigma_{N,j}^k - \sigma_{N,j}\| \lesssim d(\theta^k, \theta_0) \text{ for } j = 1, \ldots, r, \tag{A.37}$$

and it can be shown that

$$\|F^{k,r} - F_0\|/\sqrt{T} \lesssim d(\theta^k, \theta_0), \quad \|\Lambda^{k,r} - \Lambda_0\|/\sqrt{N} \lesssim d(\theta^k, \theta_0). \tag{A.38}$$

From (A.38),

$$\|R_T^{-r}\| = \|F^{k,-r'} F_0/T\| = \|F^{k,-r'}(F_0 - F^{k,r})/T\| \leq \frac{\|F^{k,-r'}\|}{\sqrt{T}} \cdot \frac{\|F_0 - F^{k,r}\|}{\sqrt{T}} \lesssim d(\theta^k, \theta_0). \tag{A.39}$$

Then from (A.32) and (A.39)

$$\|\Lambda^{k,-r}\|/\sqrt{N} \leq \|\Lambda^{k,-r} - \Lambda_0 R_T^{-r'}\|/\sqrt{N} + \|\Lambda_0 R_T^{-r'}\|/\sqrt{N} \lesssim d(\theta^k, \theta_0). \tag{A.40}$$

Therefore, the desired results follow from (A.38) and (A.40). □

Write

$$\mathbb{M}_{NT}(\theta^k) = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \rho_\tau(X_{it} - \lambda_i^{k'} f_t^k), \quad \bar{\mathbb{M}}_{NT}(\theta^k) = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \mathbb{E}[\rho_\tau(X_{it} - \lambda_i^{k'} f_t^k)],$$

$$\mathbb{W}_{NT}(\theta^k) = \mathbb{M}_{NT}(\theta^k) - \bar{\mathbb{M}}_{NT}(\theta^k) - \left(\mathbb{M}_{NT}(\theta_0) - \bar{\mathbb{M}}_{NT}(\theta_0)\right).$$

Lemma 5. *Suppose that Assumption 1 holds and $r < k < \infty$. For sufficiently small δ, we have:*

$$\mathbb{E}\left[\sup_{\theta^k \in \Theta^k(\delta)} \left|\mathbb{W}_{NT}(\theta^k)\right|\right] \lesssim \frac{\delta}{L_{NT}}.$$

Proof. Similar to the proof of Lemma 3, we can show that

$$\mathbb{E}\left[\sup_{\Theta^k(\delta)} \sqrt{NT}\left|\mathbb{W}_{NT}(\theta^k)\right|\right] \lesssim \int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^k(\delta))} d\epsilon. \tag{A.41}$$

Thus, it remains to prove that

$$\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^k(\delta))} d\epsilon = O((N+T)\delta). \tag{A.42}$$

To show (A.42), note that according to the previous lemma, $d(\theta^k,\theta_0)\le\delta$ implies that

$$\|F^{k,r}-F_0\|/\sqrt{T}+\|\Lambda^{k,r}-\Lambda_0\|/\sqrt{N}+\|\Lambda^{k,-r}\|/\sqrt{N}\le K_{10}\delta.$$

Moreover, we have for some $K_{11}>0$,

$$\|F^{k,r}-F_0\|/\sqrt{T}+\|\Lambda^{k,r}-\Lambda_0\|/\sqrt{N}+\|\Lambda^{k,-r}\|/\sqrt{N}\ge K_{11}\sqrt{\frac{\|F^{k,r}-F_0\|^2}{T}+\frac{\|\Lambda^{k,r}-\Lambda_0\|^2}{N}+\frac{\|\Lambda^{k,-r}\|^2}{N}}.$$

Thus, the set $\Theta^k(\delta)$ is contained in $\Theta^{k**}(\delta)$ where

$$\Theta^{k**}(\delta)=\left\{\theta^k\in\Theta^k:\sqrt{\frac{\|F^{k,r}-F_0\|^2}{T}+\frac{\|\Lambda^{k,r}-\Lambda_0\|^2}{N}+\frac{\|\Lambda^{k,-r}\|^2}{N}}\le K_{12}\delta,\frac{\|F^{k,-r}\|}{\sqrt{T}}\le K_{13}\right\}.$$

In addition, similar to the proof of Lemma 3, we can show that for $\theta_a^k,\theta_b^k\in\Theta^k(\delta)$, $d(\theta_a^k,\theta_b^k)\le d^{**}(\theta_a^k,\theta_b^k)$, where

$$d^{**}(\theta_a^k,\theta_b^k)=K_{14}\sqrt{\frac{\|F_a^{k,r}-F_b^{k,r}\|^2}{T}+\frac{\|\Lambda_a^{k,r}-\Lambda_b^{k,r}\|^2+\|\Lambda_a^{k,-r}-\Lambda_b^{k,-r}\|^2}{N}}+K_{15}\delta\cdot\frac{\|F_a^{k,-r}-F_b^{k,-r}\|}{\sqrt{T}}.$$

Then, we have $D(\epsilon,d,\Theta^k(\delta))\le D(\epsilon,d^{**},\Theta^{k**}(\delta))$.

Next, we calculate $D(\epsilon,d^{**},\Theta^{k**}(\delta))$. Let $(F_1^{k,r},\Lambda_1^k),\ldots,(F_{m(\epsilon)}^{k,r},\Lambda_{m(\epsilon)}^k)$ be a maximal set of points in $\Theta^{k**}(\delta)$ such that

$$\sqrt{\frac{\|F_p^{k,r}-F_q^{k,r}\|^2}{T}+\frac{\|\Lambda_p^{k,r}-\Lambda_q^{k,r}\|^2}{N}+\frac{\|\Lambda_p^{k,-r}-\Lambda_q^{k,-r}\|^2}{N}}>\epsilon/(2K_{14})$$

for any $p,q\le m(\epsilon)$ and $p\ne q$. Similar to the proof of Lemma 3, we can show that $m(\epsilon)=(K_{16}\delta/\epsilon)^{Tr+Nk}$. Let $F_1^{k,-r},\ldots,F_{n(\epsilon)}^{k,-r}$ be a maximal set of points in $\Theta^{k**}(\delta)$ such that

$$\frac{\|F_p^{k,-r}-F_q^{k,-r}\|}{\sqrt{T}}>\epsilon/(2K_{15}\delta)$$

for any $p,q\le n(\epsilon)$ and $p\ne q$. Then similarly it can be shown that $n(\epsilon)=(K_{17}\delta/\epsilon)^{T(k-r)}$. Therefore, for any $\theta^k\in\Theta^{k**}(\delta)$, we can find $p^*\le m(\epsilon)$ and $q^*\le n(\epsilon)$ such that

$$\sqrt{\frac{\|F^{k,r}-F_{p^*}^{k,r}\|^2}{T}+\frac{\|\Lambda^{k,r}-\Lambda_{p^*}^{k,r}\|^2}{N}+\frac{\|\Lambda^{k,-r}-\Lambda_{p^*}^{k,-r}\|^2}{N}}\le\epsilon/(2K_{14}),\frac{\|F^{k,-r}-F_{q^*}^{k,-r}\|}{\sqrt{T}}\le\epsilon/(2K_{15}\delta).$$

Let θ^{k*} consist of $F_{p^*}^{k,r},F_{q^*}^{k,-r},\Lambda_{p^*}^{k,r}$ and $\Lambda_{p^*}^{k,-r}$. Then it follows by the definition of d^{**} that $d^{**}(\theta^k,\theta^{k*})\le\epsilon$. As θ^k varies in $\Theta^{k**}(\delta)$, the number of possible choices for θ^{k*} is bounded by $m(\epsilon)\cdot n(\epsilon)=(K_{18}\delta/\epsilon)^{(T+N)k}$. This means that

$$D(\epsilon,d,\Theta^k(\delta))\le D(\epsilon,d^{**},\Theta^{k**}(\delta))\le(K_{18}\delta/\epsilon)^{(T+N)k}.\tag{A.43}$$

Finally, (A.42) is easily obtained from (A.43) and the desired result follows. □

Proof of Theorem 2:

Proof. First, similar to the proof of Lemma 1, we can show that $d(\hat{\theta}^k, \theta_0) = o_P(1)$. Second, similar to the proof of Theorem 1, it follows from the previous lemma that

$$d(\hat{\theta}^k, \theta_0) = O_P(L_{NT}^{-1}). \tag{A.44}$$

Next, from (A.37), (A.40) and Assumption 1

$$|\hat{\sigma}_{N,j}^k - \sigma_j| = o_P(1) \text{ for } j = 1, \ldots, r, \tag{A.45}$$

and

$$\sum_{j=r+1}^{k} \hat{\sigma}_{N,j}^k = \|\hat{\Lambda}^{k,-r}\|^2/N \leq d(\hat{\theta}^k, \theta_0)^2 = O_P(L_{NT}^{-2}). \tag{A.46}$$

Thus, by (A.45) and (A.46), we have

$$P[\hat{r}_{\text{rank}} \neq r] = P[\hat{r}_{\text{rank}} < r] + P[\hat{r}_{\text{rank}} > r] \leq P[\hat{\sigma}_{N,r}^k \leq P_{NT}] + P[\hat{\sigma}_{N,r+1}^k > P_{NT}] = o(1). \tag{A.47}$$

Then it follows that $P[\hat{r}_{\text{rank}} = r] \to 1$. □

A.3 Proof of Theorem 3

Proof. Following the proof of Bai and Ng (2002), it suffices to show that for some $C > 0$,

$$\mathbb{M}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\hat{\theta}^r) > C + o_P(1) \text{ for } l < r, \tag{A.48}$$

and

$$\mathbb{M}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\hat{\theta}^r) = o_P(1/L_{NT}^2) \text{ for } l > r. \tag{A.49}$$

Adding and subtracting terms we can write

$$\begin{aligned}\mathbb{M}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\hat{\theta}^r) = &\Big(\mathbb{M}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\theta_0) + \bar{\mathbb{M}}_{NT}(\theta_0)\Big) \\ &- \Big(\mathbb{M}_{NT}(\hat{\theta}^r) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^r) - \mathbb{M}_{NT}(\theta_0) + \bar{\mathbb{M}}_{NT}(\theta_0)\Big) + \bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^r). \end{aligned} \tag{A.50}$$

Case 1: Consider $l < r$.

Let K denote a generic positive constant. Similar to the proof of Lemma 1, it can be shown that the first two terms on the RHS of (A.50) are both $o_P(1)$, and for the last term we have $\bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^r) \geq K d^2(\hat{\theta}^l, \hat{\theta}^r) + o_P(1)$. Next, similar to (A.9) we can show that $\|M_{\hat{F}^l}\hat{F}^r\|/\sqrt{T} \lesssim d(\hat{\theta}^l, \hat{\theta}^r)$. It then follows that

$$\bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^r) \geq K\|M_{\hat{F}^l}\hat{F}^r\|^2/T + o_P(1). \tag{A.51}$$

Note that

$$\|M_{\hat{F}^l}\hat{F}^r\|^2/T = \text{Trace}\left[\mathbb{I}_r - \hat{F}^{r'}\hat{F}^l\hat{F}^{l'}\hat{F}^r/T^2\right] \geq \rho_{max}\left[\mathbb{I}_r - \hat{F}^{r'}\hat{F}^l\hat{F}^{l'}\hat{F}^r/T^2\right]. \tag{A.52}$$

By Lemma A.5 of Ahn and Horenstein (2013),

$$\rho_{max}\left[\mathbb{I}_r - \hat{F}^{r\prime}\hat{F}^l\hat{F}^{l\prime}\hat{F}^r/T^2\right] + \rho_{min}\left[\hat{F}^{r\prime}\hat{F}^l\hat{F}^{l\prime}\hat{F}^r/T^2\right] \geq \rho_{min}\left[\mathbb{I}_r\right]. \tag{A.53}$$

Since $\hat{F}^{r\prime}\hat{F}^l\hat{F}^{l\prime}\hat{F}^r$ is a $r \times r$ symmetric matrix with rank less or equal to l, we have $\rho_{min}\left[\hat{F}^{r\prime}\hat{F}^l\hat{F}^{l\prime}\hat{F}^r/T^2\right] = 0$, and the above inequality implies that

$$\rho_{max}\left[\mathbb{I}_r - \hat{F}^{r\prime}\hat{F}^l\hat{F}^{l\prime}\hat{F}^r/T^2\right] \geq 1. \tag{A.54}$$

Thus, (A.48) follows from (A.50) to (A.54).

Case 2: Now consider $l > r$.

First, similar to the proof of Theorem 2, it can be shown that for sufficiently small δ,

$$\mathbb{E}\left[\sup_{d(\theta^l,\theta_0)\leq\delta}\left|\mathbb{M}_{NT}(\theta^l) - \bar{\mathbb{M}}_{NT}(\theta^l) - \mathbb{M}_{NT}(\theta_0) + \bar{\mathbb{M}}_{NT}(\theta_0)\right|\right] \lesssim \frac{\delta}{L_{NT}},$$

and $d(\hat{\theta}^l, \theta_0) = O_P(1/L_{NT})$. It then follows that

$$\mathbb{M}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\theta_0) + \bar{\mathbb{M}}_{NT}(\theta_0) = O_P(1/L_{NT}^2). \tag{A.55}$$

Second, similar to the proof of Lemma 3 and Theorem 1 we can show that

$$\mathbb{M}_{NT}(\hat{\theta}^r) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^r) - \mathbb{M}_{NT}(\theta_0) + \bar{\mathbb{M}}_{NT}(\theta_0) = O_P(1/L_{NT}^2). \tag{A.56}$$

Finally, consider $\bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^r)$. We can write

$$\bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^r) = \bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\theta_0) - \left(\bar{\mathbb{M}}_{NT}(\hat{\theta}^r) - \bar{\mathbb{M}}_{NT}(\theta_0)\right).$$

Similarly to the proof of Lemma 1, we can show that

$$\bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\theta_0) \lesssim d^2(\hat{\theta}^l, \theta_0) \quad \text{and} \quad \bar{\mathbb{M}}_{NT}(\hat{\theta}^r) - \bar{\mathbb{M}}_{NT}(\theta_0) \lesssim d^2(\hat{\theta}^r, \theta_0).$$

It then follows from $d(\hat{\theta}^l, \theta_0) = O_P(1/L_{NT})$ and $d(\hat{\theta}^r, \theta_0) = O_P(1/L_{NT})$ that

$$\bar{\mathbb{M}}_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}_{NT}(\hat{\theta}^r) = O_P(1/L_{NT}^2). \tag{A.57}$$

Thus, (A.49) follows from (A.50), (A.55), (A.56) and (A.57). Then, this concludes the proof. □

A.4 Proof of Theorem 4

We only prove the asymptotic distribution of $\tilde{\lambda}_i$ since the proof for $\tilde{f}_t$ is symmetric. Define $\varrho(u) = [\tau - K(u/h)]u$, then we can write

$$\mathbb{S}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\varrho(X_{it} - \lambda_i' f_t).$$

Let $\varrho^{(j)}(u) = (\partial/\partial u)^j \varrho(u)$ for $j = 1, 2, 3$. For fixed λ_i, f_t, define

$$\bar{\varrho}(X_{it} - \lambda_i' f_t) = \mathbb{E}\left[\varrho(X_{it} - \lambda_i' f_t)\right], \quad \bar{\varrho}^{(j)}(X_{it} - \lambda_i' f_t) = \mathbb{E}\left[\varrho^{(j)}(X_{it} - \lambda_i' f_t)\right] \text{ for } j = 1, 2, 3.$$

When the functions defined above are evaluated at the true parameters, we suppress their arguments to further simplify the notations. For example, $\varrho_{it} = \varrho(X_{it} - \lambda_{0i}' f_{0t})$, $\bar{\varrho}_{it} = \bar{\varrho}(X_{it} - \lambda_{0i}' f_{0t})$. Moreover, define

$$\bar{\mathbb{S}}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\bar{\varrho}(X_{it} - \lambda_i' f_t),$$

$$\mathbb{U}_{NT}(\theta) = \mathbb{S}_{NT}(\theta) - \bar{\mathbb{S}}_{NT}(\theta) - (\mathbb{S}_{NT}(\theta_0) - \bar{\mathbb{S}}_{NT}(\theta_0)).$$

Using $\bar{O}(1)$ to denote a sequence that is uniformly (over i and t) bounded.

Lemma 6. *Under Assumptions 1 and 2,*
(i) There exists a constant $\bar{C} > 0$ such that $h^{j-1}|\varrho^{(j)}(u)| \leq \bar{C}$ for $j = 1, 2, 3$.
(ii) $\bar{\varrho}_{it}^{(1)} = \bar{O}(h^m)$, $\bar{\varrho}^{(2)}(X_{it} - \lambda_i' f_t) = \mathsf{f}_{it}(\lambda_i' f_t - \lambda_{0i}' f_{0t}) + \bar{O}(h^m)$, and $\bar{\varrho}^{(3)}(X_{it} - \lambda_i' f_t) = \mathsf{f}_{it}^{(1)}(\lambda_i' f_t - \lambda_{0i}' f_{0t}) + \bar{O}(h^m)$.
(iii) $\mathbb{E}(\varrho_{it}^{(1)})^2 = \tau(1-\tau) + \bar{O}(h)$, and $h \cdot \mathbb{E}\left[(\varrho_{it}^{(2)})^2\right] = \bar{O}(1)$.

Proof. The proof is similar to the standard calculations of the means of kernel density estimators, and it is omitted here to save space. Similar results can be found in Horowitz (1998) and Galvao and Kato (2016). □

Lemma 7. *Under Assumptions 1 and 2, $d(\tilde{\theta}, \theta_0) = o_P(1)$ as $N, T \to \infty$ and $h \to 0$.*

Proof. By definition we have $\mathbb{S}_{NT}(\tilde{\theta}) \leq \mathbb{S}_{NT}(\theta_0)$. Adding and subtracting terms and using (A.1) we have

$$d^2(\tilde{\theta}, \theta_0) \lesssim \bar{\mathbb{M}}_{NT}(\tilde{\theta}) - \bar{\mathbb{M}}_{NT}(\theta_0) \leq \mathbb{M}_{NT}(\tilde{\theta}) - \mathbb{S}_{NT}(\tilde{\theta}) + \mathbb{S}_{NT}(\theta_0) - \mathbb{M}_{NT}(\theta_0) + \bar{\mathbb{M}}_{NT}(\tilde{\theta}) - \mathbb{M}_{NT}(\tilde{\theta}) + \mathbb{M}_{NT}(\theta_0) - \bar{\mathbb{M}}_{NT}(\theta_0).$$

It follows that

$$d^2(\tilde{\theta}, \theta_0) \lesssim \sup_{\theta \in \Theta^M} |\mathbb{M}_{NT}(\theta) - \mathbb{S}_{NT}(\theta)| + \sup_{\theta \in \Theta^M} |\mathbb{W}_{NT}(\theta)|.$$

It is easy to see that the first term on the RHS of the above inequality is $O(h)$, and the second term is $o_P(1)$ as proved in Lemma 1. Then the desired result follows. □

Lemma 8. *Under Assumptions 1 and 2, $d(\tilde{\theta}, \theta_0) = O_P(1/L_{NT})$ as $N, T \to \infty$.*

Proof. First, since $\varrho^{(1)}(u)$ is uniformly bounded, we have $|\varrho(X_{it} - c_1) - \varrho(X_{it} - c_2)| \lesssim |c_1 - c_2|$. Then similar to the proof of Lemma 3 it can be shown that

$$\mathbb{E}\left[\sup_{\theta \in \Theta^M(\delta)} |\mathbb{U}_{NT}(\theta)|\right] \lesssim \frac{\delta}{L_{NT}}. \tag{A.58}$$

Similar to the proof Theorem 1, the parameter space Θ^M can be partitioned into shells $S_j = \{\theta \in \Theta^M : 2^{j-1} < L_{NT} \cdot d(\theta, \theta_0) \leq 2^j\}$. Concluding that, for a given integer V and for every $\eta > 0$,

$$P\left[L_{NT} \cdot d(\tilde{\theta}, \theta_0) > 2^V\right] \leq \sum_{j \geq V, 2^j \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} (\mathbb{S}_{NT}(\theta) - \mathbb{S}_{NT}(\theta_0)) \leq 0\right] + P[d(\tilde{\theta}, \theta_0) \geq \eta].$$

For arbitrarily small $\eta > 0$, the second probability on the RHS of the above equation converges to 0 as $N, T \to \infty$ by Lemma 7.

Next, expanding $\mathbb{S}_{NT}(\theta)$ around θ_0 and taking expectations

$$\bar{\mathbb{S}}_{NT}(\theta) - \bar{\mathbb{S}}_{NT}(\theta_0) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \bar{\varrho}_{it}^{(1)} \cdot (\lambda_i' f_t - \lambda_{0i}' f_{0t}) + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \bar{\varrho}^{(2)}(c_{it}^*) \cdot (\lambda_i' f_t - \lambda_{0i}' f_{0t})^2,$$

where c_{it}^* lies between $\lambda_i' f_t$ and $\lambda_{0i}' f_{0t}$. Then, it follows from Lemma 6 and Assumption 2 that

$$\bar{\mathbb{S}}_{NT}(\theta) - \bar{\mathbb{S}}_{NT}(\theta_0) \geq O(h^m) + \underline{\mathsf{f}} \cdot d^2(\theta, \theta_0).$$

Thus, for each θ in S_j we have

$$-[\bar{\mathbb{S}}_{NT}(\theta) - \bar{\mathbb{S}}_{NT}(\theta_0)] \leq -\underline{\mathsf{f}} \cdot d_{NT}^2(\theta, \theta_0) + O(h^m) \leq -\underline{\mathsf{f}} \cdot \frac{2^{2j-2}}{L_{NT}^2} + O(h^m).$$

Therefore, $\inf_{\theta \in S_j}(\mathbb{S}_{NT}(\theta) - \mathbb{S}_{NT}(\theta_0)) \leq 0$ implies that

$$\inf_{\theta \in S_j} \mathbb{U}_{NT}(\theta) \leq -\underline{\mathsf{f}} \cdot \frac{2^{2j-2}}{L_{NT}^2} + O(h^m),$$

and it follows that

$$\sum_{j \geq V, 2^j \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} (\mathbb{S}_{NT}(\theta) - \mathbb{S}_{NT}(\theta_0)) \leq 0\right] \leq \sum_{j \geq V, 2^j \leq \eta L_{NT}} P\left[\sup_{\theta \in S_j} |\mathbb{U}_{NT}(\theta)| \geq \underline{\mathsf{f}} \cdot \frac{2^{2j-2}}{L_{NT}^2} + O(h^m)\right].$$

By (A.58) and Markov's inequality,

$$P\left[\sup_{\theta \in S_j} |\mathbb{U}_{NT}(\theta)| \geq \underline{\mathsf{f}} \cdot \frac{2^{2j-2}}{L_{NT}^2} + O(h^m)\right] \lesssim \frac{2^j}{2^{2j} + O(L_{NT}^2 \cdot h^m)}.$$

By Assumption 2, $O(L_{NT}^2 \cdot h^m) = o(1)$. Thus, the above inequality implies that

$$\sum_{j \geq V, 2^j \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} (\mathbb{M}_{NT}(\theta) - \mathbb{M}_{NT}(\theta_0)) \leq 0\right] \lesssim \sum_{j \geq V} 2^{-j}.$$

The RHS of the previous expression convergences to 0 as $V \to \infty$, implying that $L_{NT} \cdot d(\tilde{\theta}, \theta_0) = O_P(1)$, or $d(\tilde{\theta}, \theta_0) = O_P(1/L_{NT})$. □

Define:

$$\mathbb{S}_{i,T}(\lambda, F) = \frac{1}{T}\sum_{t=1}^{T} \varrho(X_{it} - \lambda' f_t), \quad \bar{\mathbb{S}}_{i,T}(\lambda, F) = \frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}(X_{it} - \lambda' f_t),$$

and

$$\mathbb{M}_{i,T}(\lambda, F) = \frac{1}{T}\sum_{t=1}^{T} \rho_\tau(X_{it} - \lambda' f_t), \quad \bar{\mathbb{M}}_{i,T}(\lambda, F) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\rho_\tau(X_{it} - \lambda' f_t).$$

Lemma 9. *Under Assumptions 1 and 2, $\|\tilde{\lambda}_i - \lambda_{0i}\| = o_P(1)$ for each i.*

Proof. Note that

$$\tilde{\lambda}_i = \arg\min_{\lambda \in \mathcal{A}} \mathbb{S}_{i,T}(\lambda, \tilde{F}).$$

First, we show that

$$\sup_{\lambda \in \mathcal{A}} |\mathbb{S}_{i,T}(\lambda, \tilde{F}) - \bar{\mathbb{M}}_{i,T}(\lambda, F_0)| = o_P(1). \tag{A.59}$$

Adding and subtracting terms we have

$$\mathbb{S}_{i,T}(\lambda, \tilde{F}) - \bar{\mathbb{M}}_{i,T}(\lambda, F_0) = \mathbb{S}_{i,T}(\lambda, \tilde{F}) - \mathbb{M}_{i,T}(\lambda, \tilde{F}) + \mathbb{M}_{i,T}(\lambda, \tilde{F}) - \mathbb{M}_{i,T}(\lambda, F_0) + \mathbb{M}_{i,T}(\lambda, F_0) - \bar{\mathbb{M}}_{i,T}(\lambda, F_0).$$

Then,

$$\begin{aligned}\sup_{\lambda \in \mathcal{A}} |\mathbb{S}_{i,T}(\lambda, \tilde{F}) - \bar{\mathbb{M}}_{i,T}(\lambda, F_0)| \leq & \sup_{\lambda \in \mathcal{A}} |\mathbb{S}_{i,T}(\lambda, \tilde{F}) - \mathbb{M}_{i,T}(\lambda, \tilde{F})| + \\ & \sup_{\lambda \in \mathcal{A}} |\mathbb{M}_{i,T}(\lambda, \tilde{F}) - \mathbb{M}_{i,T}(\lambda, F_0)| + \sup_{\lambda \in \mathcal{A}} |\mathbb{M}_{i,T}(\lambda, F_0) - \bar{\mathbb{M}}_{i,T}(\lambda, F_0)|.\end{aligned}$$

It is easy to show that

$$\sup_{\lambda \in \mathcal{A}} |\mathbb{S}_{i,T}(\lambda, \tilde{F}) - \mathbb{M}_{i,T}(\lambda, \tilde{F})| \lesssim h,$$

$$\sup_{\lambda \in \mathcal{A}} |\mathbb{M}_{i,T}(\lambda, \tilde{F}) - \mathbb{M}_{i,T}(\lambda, F_0)| \lesssim \sup_{\lambda \in \mathcal{A}} \|\lambda\| \cdot \frac{1}{T}\sum_{t=1}^{T} \|\tilde{f}_t - f_{0t}\| \lesssim \|\tilde{F} - F_0\|/\sqrt{T} = O_P(1/L_{NT}),$$

$$\sup_{\lambda \in \mathcal{A}} |\mathbb{M}_{i,T}(\lambda, F_0) - \bar{\mathbb{M}}_{i,T}(\lambda, F_0)| = \sup_{\lambda \in \mathcal{A}} \left|\frac{1}{T}\sum_{t=1}^{T} [\rho_\tau(X_{it} - \lambda' f_{0t}) - \mathbb{E}\rho_\tau(X_{it} - \lambda' f_{0t})]\right| = o_P(1).$$

Then (A.59) follows as $h \to 0$.

Second, it can be shown that for any $\epsilon > 0$, and $B_i(\epsilon) = \{\lambda \in \mathcal{A} : \|\lambda - \lambda_{0i}\| \leq \epsilon\}$,

$$\inf_{\lambda \in B_i^C(\epsilon)} \bar{\mathbb{M}}_{i,T}(\lambda, F_0) - \bar{\mathbb{M}}_{i,T}(\lambda_{0i}, F_0) > 0, \tag{A.60}$$

e.g., the proof of Proposition 3.1 of Galvao and Kato (2016).

Finally, given (A.59) and (A.60), the consistency of $\tilde{\lambda}_i$ follows from a standard proof for the consistency of M-estimators (see Theorem 2.1 of Newey and McFadden 1994). $\square$

Lemma 10. *Under Assumptions 1 and 2,* $\|\tilde{\lambda}_i - \lambda_{0i}\| = O_P\left(T^{-1/2}h^{-1}\right)$ *for each* i.

Proof. For any fixed $\lambda_i \in \mathcal{A}$ and $f_t \in \mathcal{F}$, expanding $\varrho^{(1)}(X_{it} - \lambda_i' f_t) f_t$ gives

$$\begin{aligned}
&\varrho^{(1)}(X_{it} - \lambda_i' f_t) f_t \\
=&\varrho^{(1)}(X_{it} - \lambda_{0i}' f_t) f_t - \varrho^{(2)}(X_{it} - \lambda_{0i}' f_t) f_t f_t' \cdot (\lambda_i - \lambda_{0i}) + 0.5\varrho^{(3)}(X_{it} - \lambda_i^{*\prime} f_t) f_t [(\lambda_i - \lambda_{0i})' f_t]^2 \\
=&\varrho_{it}^{(1)} f_{0t} + \varrho^{(1)}(X_{it} - \lambda_{0i}' f_t^*)(f_t - f_{0t}) - \varrho^{(2)}(X_{it} - \lambda_{0i}' f_t^*) f_t^* \lambda_{0i}'(f_t - f_{0t}) - \varrho_{it}^{(2)} f_t f_t' \cdot (\lambda_i - \lambda_{0i}) \\
&+ \varrho^{(3)}(X_{it} - \lambda_{0i}' f_t^*) f_t f_t' \cdot (\lambda_i - \lambda_{0i}) \lambda_{0i}'(f_t - f_{0t}) + 0.5\varrho^{(3)}(X_{it} - \lambda_i^{*\prime} f_t) f_t [(\lambda_i - \lambda_{0i})' f_t]^2,
\end{aligned}$$

where λ_i^* lies between λ_i and λ_{0i} and f_t^* lies between f_t and f_{0t}. Taking expectations of both sides of the above equation, and setting $\lambda_i = \tilde{\lambda}_i$, $f_t = \tilde{f}_t$, it follows from Lemma 4 that:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}^{(1)}(X_{it} - \tilde{\lambda}_i' \tilde{f}_t)\tilde{f}_t =& \frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}_{it}^{(1)} f_{0t} - \left(\frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}_{it}^{(2)} \tilde{f}_t \tilde{f}_t'\right)(\tilde{\lambda}_i - \lambda_{0i}) \\
&+ O_P\left(T^{-1/2}\|\tilde{F} - F_0\|\right) + O_P(\|\tilde{\lambda}_i - \lambda_{0i}\|) \cdot O_P\left(T^{-1/2}\|\tilde{F} - F_0\|\right) + O_P(\|\tilde{\lambda}_i - \lambda_{0i}\|^2).
\end{aligned}$$

Lemma 6, Lemma 8 and Assumption 2 imply that:

$$\frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}_{it}^{(2)} \tilde{f}_t \tilde{f}_t' = \frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}_{it}^{(2)} f_{0t} f_{0t} + o_P(1) = \Phi_i + o_P(1).$$

Then, from Lemma 6, Lemma 8 and Lemma 9 we get

$$\Phi_i(\tilde{\lambda}_i - \lambda_{0i}) + o_P(\|\tilde{\lambda}_i - \lambda_{0i}\|) = O(h^m) + O_P(1/L_{NT}) - \frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}^{(1)}(X_{it} - \tilde{\lambda}_i' \tilde{f}_t)\tilde{f}_t. \tag{A.61}$$

Note that we can write

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}^{(1)}(X_{it} - \tilde{\lambda}_i' \tilde{f}_t)\tilde{f}_t \\
=& -\frac{1}{T}\sum_{t=1}^{T} \tilde{\varrho}_{it}^{(1)} f_{0t} - \frac{1}{T}\sum_{t=1}^{T}\left[\tilde{\varrho}^{(1)}(X_{it} - \tilde{\lambda}_i' \tilde{f}_t)\tilde{f}_t - \tilde{\varrho}_{it}^{(1)} f_{0t}\right] \\
=& -\frac{1}{T}\sum_{t=1}^{T} \tilde{\varrho}_{it}^{(1)} f_{0t} - \frac{1}{T}\sum_{t=1}^{T}\left[\tilde{\varrho}^{(1)}(X_{it} - \tilde{\lambda}_i' \tilde{f}_t)\tilde{f}_t - \tilde{\varrho}^{(1)}(X_{it} - \tilde{\lambda}_i' f_{0t}) f_{0t}\right] - \frac{1}{T}\sum_{t=1}^{T}\left[\tilde{\varrho}^{(1)}(X_{it} - \tilde{\lambda}_i' f_{0t}) - \tilde{\varrho}_{it}^{(1)}\right] f_{0t}.
\end{aligned}$$

The first term on the RHS of the above equation is $O_P(T^{-1/2})$ by Lemma 6 and Lyapunov's CLT. For the second term on the right of the above equation, we have

$$\frac{1}{T}\sum_{t=1}^{T}\left[\tilde{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'\tilde{f}_t)\tilde{f}_t-\tilde{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'f_{0t})f_{0t}\right]=$$

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'f_t^*)(\tilde{f}_t-f_{0t})-\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}^{(2)}(X_{it}-\tilde{\lambda}_i'f_t^*)f_t^*\tilde{\lambda}_i'(\tilde{f}_t-f_{0t}), \quad \text{(A.62)}$$

where f_t^* lies between $\tilde{f}_t$ and f_{0t}. The first term on the right of (A.62) is $O_P(1/L_{NT})$ because $\varrho^{(1)}$ is uniformly bounded and $T^{-1}\sum_{t=1}^{T}\|\tilde{f}_t-f_{0t}\|=O_P(1/L_{NT})$ by Lemma 8. Similarly, the second term on the RHS of (A.62) is $O_P(1/(L_{NT}h))$ because $h\varrho^{(2)}(u)$ is uniformly bounded. Finally, we can show that (see, e.g., Lemma B.2 of Galvao and Kato 2016)

$$\frac{1}{T}\sum_{t=1}^{T}\left[\tilde{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'f_{0t})-\tilde{\varrho}_{it}^{(1)}\right]f_{0t}=O_P(\|\tilde{\lambda}_i-\lambda_{0i}\|)\cdot O_P(1/\sqrt{Th})=o_P(\|\tilde{\lambda}_i-\lambda_{0i}\|).$$

Combining the above results we have

$$\frac{1}{T}\sum_{t=1}^{T}\bar{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'\tilde{f}_t)\tilde{f}_t=O_P\left(\frac{1}{L_{NT}h}\right)+o_P(\|\tilde{\lambda}_i-\lambda_{0i}\|), \quad \text{(A.63)}$$

and the desired result follows from (A.61), (A.63) and Assumption 2. □

To derive the asymptotic distribution of $\tilde{\lambda}_i$, it is essential to obtain the stochastic expansion of $\tilde{f}_t$. Define

$$\mathbb{P}_{NT}(\theta)=b\left[\frac{1}{2N}\sum_{p=1}^{r}\sum_{q>p}^{r}\left(\sum_{i=1}^{N}\lambda_{ip}\lambda_{iq}\right)^2+\frac{1}{2T}\sum_{p=1}^{r}\sum_{q>p}^{r}\left(\sum_{t=1}^{T}f_{tp}f_{tq}\right)^2+\frac{1}{8T}\sum_{k=1}^{r}\left(\sum_{t=1}^{T}f_{tk}^2-T\right)^2\right]$$

for some $b>0$. Define

$$\mathcal{S}^*(\theta)=\Big[\underbrace{\ldots,-\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\bar{\varrho}^{(1)}(X_{it}-\lambda_i'f_t)f_t',\ldots}_{1\times Nr},\underbrace{\ldots,-\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\bar{\varrho}^{(1)}(X_{it}-\lambda_i'f_t)\lambda_i',\ldots}_{1\times Tr}\Big].$$

$$\mathcal{S}(\theta)=\mathcal{S}^*(\theta)+\partial\mathbb{P}_{NT}(\theta)/\partial\theta, \quad \mathcal{H}(\theta)=\partial\mathcal{S}^*(\theta)/\partial\theta'+\partial^2\mathbb{P}_{NT}(\theta)/\partial\theta\partial\theta',$$

and let $\mathcal{H}=\mathcal{H}(\theta_0)$. Expanding $\mathcal{S}(\tilde{\theta})$ around $\mathcal{S}(\theta_0)$ gives:

$$\mathcal{S}(\tilde{\theta})=\mathcal{S}(\theta_0)+\mathcal{H}\cdot(\tilde{\theta}-\theta_0)+0.5\mathcal{R}(\tilde{\theta}), \quad \text{(A.64)}$$

where

$$\mathcal{R}(\tilde{\theta})=\left(\sum_{j=1}^{M}\partial\mathcal{H}(\theta^*)/\partial\theta_j\cdot(\tilde{\theta}_j-\theta_{0j})\right)(\tilde{\theta}-\theta_0),$$

and θ^* lies between $\tilde{\theta}$ and θ_0.

Further, define

$$\mathcal{H}_d = \begin{pmatrix} \mathcal{H}_d^{\Lambda} & 0 \\ 0 & \mathcal{H}_d^{F} \end{pmatrix}, \quad \mathcal{H}_d^{\Lambda} = \frac{\sqrt{T}}{\sqrt{N}} \operatorname{diag}\left[\Phi_{T,1}, \ldots, \Phi_{T,i}, \ldots, \Phi_{T,N}\right], \quad \mathcal{H}_d^{F} = \frac{\sqrt{N}}{\sqrt{T}} \operatorname{diag}\left[\Psi_{N,1}, \ldots, \Psi_{N,t}, \ldots, \Psi_{N,T}\right].$$

where

$$\Phi_{T,i} = \frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}_{it}^{(2)} f_{0t} f_{0t}', \quad \Psi_{N,t} = \frac{1}{N}\sum_{i=1}^{N} \bar{\varrho}_{it}^{(2)} \lambda_{0i}\lambda_{0i}'.$$

The following lemma is important for the stochastic expansion of $\tilde{f}_t$.

Lemma 11. *Under Assumptions 1 and 2, the matrix $\mathcal{H}$ is invertible and $\|\mathcal{H}^{-1} - \mathcal{H}_d^{-1}\|_{\max} = O(1/T)$.*

Proof. To simplify the notations, we consider the case $r = 2$, but the proof can be easily generalized to the case $r > 2$. Note that $\lambda_{0i} = (\lambda_{0i,1}, \lambda_{0i,2})'$ and $f_{0t} = (f_{0t,1}, f_{0t,2})'$.

First, define

$$\begin{aligned}
\gamma_1' &= \left[\mathbf{0}_{1\times 2N}, (f_{01,1}, 0), \ldots, (f_{0t,1}, 0), \ldots, (f_{0T,1}, 0)\right]/\sqrt{T}, \\
\gamma_2' &= \left[\mathbf{0}_{1\times 2N}, (0, f_{01,2}), \ldots, (0, f_{0t,2}), \ldots, (0, f_{0T,2})\right]/\sqrt{T}, \\
\gamma_3' &= \left[\mathbf{0}_{1\times 2N}, (f_{01,2}, f_{01,1}), \ldots, (f_{0t,2}, f_{0t,1}), \ldots, (f_{0T,2}, f_{0T,1})\right]/\sqrt{T}, \\
\gamma_4' &= \left[(\lambda_{01,2}, \lambda_{01,1}), \ldots, (\lambda_{0i,2}, \lambda_{0i,1}), \ldots, (\lambda_{0N,2}, \lambda_{0N,1}), \mathbf{0}_{1\times 2T}\right]/\sqrt{N},
\end{aligned}$$

and note that $\partial^2 \mathbb{P}_{NT}(\theta_0)/\partial\theta\partial\theta' = b\left(\sum_{k=1}^{4} \gamma_k \gamma_k'\right)$.

Second, define

$$\begin{aligned}
\omega_1' &= \Big[\underbrace{(\lambda_{01,1}, 0)/\sqrt{N}, \ldots, (\lambda_{0N,1}, 0)/\sqrt{N}}_{\omega_{1\Lambda}'}, \underbrace{(-f_{01,1}, 0)/\sqrt{T}, \ldots, (-f_{0T,1}, 0)/\sqrt{T}}_{\omega_{1F}'}\Big], \\
\omega_2' &= \Big[\underbrace{(0, \lambda_{01,2})/\sqrt{N}, \ldots, (0, \lambda_{0N,2})/\sqrt{N}}_{\omega_{2\Lambda}'}, \underbrace{(0, -f_{01,2})/\sqrt{T}, \ldots, (0, -f_{0T,2})/\sqrt{T}}_{\omega_{2F}'}\Big], \\
\omega_3' &= \Big[\underbrace{(\lambda_{01,2}, 0)/\sqrt{N}, \ldots, (\lambda_{0N,2}, 0)/\sqrt{N}}_{\omega_{3\Lambda}'}, \underbrace{(0, -f_{01,1})/\sqrt{T}, \ldots, (0, -f_{0T,1})/\sqrt{T}}_{\omega_{3F}'}\Big], \\
\omega_4' &= \Big[\underbrace{(0, \lambda_{01,1})/\sqrt{N}, \ldots, (0, \lambda_{0N,1})/\sqrt{N}}_{\omega_{4\Lambda}'}, \underbrace{(-f_{01,2}, 0)/\sqrt{T}, \ldots, (-f_{0T,2}, 0)/\sqrt{T}}_{\omega_{4F}'}\Big],
\end{aligned}$$

and $\omega = [\omega_1, \omega_2, \omega_3, \omega_4]$. It is easy to check that $\omega_p'\omega_q = 0$ for $p \neq q$. Moreover, we have

$$\omega\omega' = \sum_{k=1}^{4} \omega_k \omega_k' = \begin{pmatrix} \sum_{k=1}^{4} \omega_{k\Lambda}\omega_{k\Lambda}' & -(NT)^{-1/2}\{f_{0t}\lambda_{0i}'\}_{i\le N, t\le T} \\ -(NT)^{-1/2}\{\lambda_{0i} f_{0t}'\}_{t\le T, i\le N,} & \sum_{k=1}^{4} \omega_{kF}\omega_{kF}' \end{pmatrix}, \tag{A.65}$$

where $\{f_{0t}\lambda_{0i}'\}_{i\le N, t\le T}$ denotes a $2N \times 2T$ matrix whose $\{i, t\}$th block is $f_{0t}\lambda_{0i}'$. Further, it is easy to see that under our normalizations,

$$\omega'\omega = \begin{pmatrix} \sigma_{N1}+1 & 0 & 0 & 0 \\ 0 & \sigma_{N2}+1 & 0 & 0 \\ 0 & 0 & \sigma_{N2}+1 & 0 \\ 0 & 0 & 0 & \sigma_{N1}+1 \end{pmatrix}.$$

Next, we project γ_k onto ω, and write $\gamma_k = \omega\beta_k + \zeta_k$ for $k = 1, \ldots, 4$, where $\beta_k = (\omega'\omega)^{-1}\omega'\gamma_k$. In particular,

$$\beta_1 = \begin{pmatrix} -\frac{1}{\sigma_{N1}+1} \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \beta_2 = \begin{pmatrix} 0 \\ -\frac{1}{\sigma_{N2}+1} \\ 0 \\ 0 \end{pmatrix}, \quad \beta_3 = \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{\sigma_{N2}+1} \\ -\frac{1}{\sigma_{N1}+1} \end{pmatrix}, \quad \beta_4 = \begin{pmatrix} 0 \\ 0 \\ \frac{\sigma_{N2}}{\sigma_{N2}+1} \\ \frac{\sigma_{N1}}{\sigma_{N1}+1} \end{pmatrix}.$$

Define $B_N = \sum_{k=1}^4 \beta_k\beta_k'$. It is easy to show that there exists $\underline{\rho} > 0$ such that $\rho_{\min}(B_N) > \underline{\rho}$ for all large N as long as $\sigma_{N1} - \sigma_{N2}$ is bounded below by a positive constant for all large N, which is true under our assumption that $\sigma_{N1} \to \sigma_1$, $\sigma_{N2} \to \sigma_2$, and $\sigma_1 > \sigma_2$. It then follows that

$$\begin{aligned} \partial^2\mathbb{P}_{NT}(\theta_0)/\partial\theta\partial\theta' = b\left(\sum_{k=1}^4 \gamma_k\gamma_k'\right) &= b\cdot\omega\left(\sum_{k=1}^4 \beta_k\beta_k'\right)\omega' + b\left(\sum_{k=1}^4 \zeta_k\zeta_k'\right) \\ &= b\underline{\rho}\cdot\omega\omega' + b\cdot\omega(B_N - \underline{\rho}\mathbb{I}_4)\omega' + b\left(\sum_{k=1}^4 \zeta_k\zeta_k'\right). \end{aligned} \quad \text{(A.66)}$$

Now let $\underline{b} = \min\{\underline{\mathrm{f}}, b\underline{\rho}\}$. Then it follows from (A.66) that:

$$\begin{aligned} \mathcal{H} &= \partial\mathcal{S}^*(\theta_0)/\partial\theta' + \partial^2\mathbb{P}_{NT}(\theta_0)/\partial\theta\partial\theta' \\ &= \partial\mathcal{S}^*(\theta_0)/\partial\theta' + \underline{b}\cdot\omega\omega' + \underbrace{(b\underline{\rho} - \underline{b})\cdot\omega\omega'}_{\geq 0} + \underbrace{b\cdot\omega(B_N - \underline{\rho}I_4)\omega'}_{\geq 0} + \underbrace{b\left(\sum_{k=1}^4 \zeta_k\zeta_k'\right)}_{\geq 0} \\ &\geq \partial\mathcal{S}^*(\theta_0)/\partial\theta' + \underline{b}\cdot\omega\omega'. \end{aligned}$$

Moreover, we can write

$$
\begin{aligned}
&\partial \mathcal{S}^*(\theta_0)/\partial\theta' \\
= &\begin{pmatrix} (NT)^{-1/2}\operatorname{diag}\left[\left\{\sum_{t=1}^T \bar{\varrho}_{it}^{(2)} f_{0t}f_{0t}'\right\}_{i\le N}\right] & (NT)^{-1/2}\left\{\bar{\varrho}_{it}^{(2)} f_{0t}\lambda_{0i}'\right\}_{i\le N, t\le T} \\ (NT)^{-1/2}\left\{\bar{\varrho}_{it}^{(2)}\lambda_{0i}f_{0t}'\right\}_{t\le T, i\le N} & (NT)^{-1/2}\operatorname{diag}\left[\left\{\sum_{i=1}^N \bar{\varrho}_{it}^{(2)}\lambda_{0i}\lambda_{0i}'\right\}_{t\le N}\right] \end{pmatrix} \\
= &\underbrace{\underline{b}\begin{pmatrix} \operatorname{diag}\left[\left\{(NT)^{-1/2}\sum_{t=1}^T f_{0t}f_{0t}'\right\}_{i\le N}\right] & \mathbf{0}_{2N\times 2T} \\ \mathbf{0}_{2T\times 2N} & \operatorname{diag}\left[\left\{(NT)^{-1/2}\sum_{i=1}^N \lambda_{0i}\lambda_{0i}'\right\}_{t\le N}\right] \end{pmatrix}}_{I} \\
&+\underbrace{\underline{b}\begin{pmatrix} \mathbf{0}_{2N\times 2N} & (NT)^{-1/2}\left\{f_{0t}\lambda_{0i}'\right\}_{i\le N, t\le T} \\ (NT)^{-1/2}\left\{\lambda_{0i}f_{0t}'\right\}_{t\le T, i\le N} & \mathbf{0}_{2T\times 2T} \end{pmatrix}}_{II} \\
&+\underbrace{\begin{pmatrix} (NT)^{-1/2}\operatorname{diag}\left[\left\{\sum_{t=1}^T (\bar{\varrho}_{it}^{(2)}-\underline{b}) f_{0t}f_{0t}'\right\}_{i\le N}\right] & (NT)^{-1/2}\left\{(\bar{\varrho}_{it}^{(2)}-\underline{b}) f_{0t}\lambda_{0i}'\right\}_{i\le N, t\le T} \\ (NT)^{-1/2}\left\{(\bar{\varrho}_{it}^{(2)}-\underline{b})\lambda_{0i}f_{0t}'\right\}_{t\le T, i\le N} & (NT)^{-1/2}\operatorname{diag}\left[\left\{\sum_{i=1}^N (\bar{\varrho}_{it}^{(2)}-\underline{b})\lambda_{0i}\lambda_{0i}'\right\}_{t\le N}\right] \end{pmatrix}}_{III}.
\end{aligned}
$$

Note that by our assumptions there exists a constant $\underline{c} > 0$ such that:

$$
I = \underline{b}\begin{pmatrix} \sqrt{T/N}\cdot \mathbb{I}_{2N} & \mathbf{0}_{2N\times 2T} \\ \mathbf{0}_{2T\times 2N} & \sqrt{N/T}\cdot \mathbb{I}_T \otimes \operatorname{diag}(\sigma_{N1},\sigma_{N2}) \end{pmatrix} \ge \underline{c}\cdot \mathbb{I}_{2(N+T)}. \tag{A.67}
$$

From (A.65) we have

$$
II + \underline{b}\cdot\omega\omega' = \underline{b}\cdot\begin{pmatrix} \sum_{k=1}^4 \omega_{k\Lambda}\omega_{k\Lambda}' & \mathbf{0}_{2N\times 2T} \\ \mathbf{0}_{2T\times 2N} & \sum_{k=1}^4 \omega_{kF}\omega_{kF}' \end{pmatrix} \ge 0. \tag{A.68}
$$

For the last term we have for N,T large enough,

$$
III = \frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T (\bar{\varrho}_{it}^{(2)} - \underline{b})\mu_{it}\mu_{it}' \ge 0, \tag{A.69}
$$

where $\mu_{it} = [\underbrace{\mathbf{0}_{1\times 2},\ldots,f_{0t}',\ldots,\mathbf{0}_{1\times 2}}_{1\times 2N},\underbrace{\mathbf{0}_{1\times 2},\ldots,\lambda_{0i}',\ldots,\mathbf{0}_{1\times 2}}_{1\times 2T}]'$, because Assumption 1 and Lemma 6 imply that $\bar{\varrho}_{it}^{(2)} \ge \underline{\mathrm{f}}$ for all i,t. It then follows from (A.67), (A.68) and (A.69) that

$$
\mathcal{H} \ge \partial\mathcal{S}^*(\theta_0)/\partial\theta' + \underline{b}\cdot\omega\omega' = I + II + III + \underline{b}\cdot\omega\omega' \ge \underline{c}\cdot\mathbb{I}_{2(N+T)},
$$

and thus

$$
\mathcal{H}^{-1} \le \underline{c}^{-1}\cdot\mathbb{I}_{2(N+T)}. \tag{A.70}
$$

Finally, write $\mathcal{H} = \mathcal{H}_d + \mathcal{C}$, where

$$\mathcal{C} = \begin{pmatrix} \mathbf{0}_{2N\times 2N} & (NT)^{-1/2}\left\{\bar{\varrho}_{it}^{(2)} f_{0t}\lambda_{0i}'\right\}_{i\le N, t\le T} \\ (NT)^{-1/2}\left\{\bar{\varrho}_{it}^{(2)} \lambda_{0i} f_{0t}'\right\}_{t\le T, i\le N} & \mathbf{0}_{2T\times 2T} \end{pmatrix} + b\left(\sum_{k=1}^{4}\gamma_k\gamma_k'\right).$$

Inequality (A.70) implies that (see Lemma 2 of Chen et al. 2018)

$$\left\|\mathcal{H}^{-1} - \mathcal{H}_d^{-1}\right\|_{\max} \le \left\|\mathcal{H}_d^{-1}\mathcal{C}\mathcal{H}_d^{-1}\right\|_{\max} + \underline{c}^{-1}\left\|\mathcal{H}_d^{-1}\mathcal{C}^2\mathcal{H}_d^{-1}\right\|_{\max}.$$

Since $\mathcal{H}_d^{-1}$ is a block-diagonal matrix whose elements are all $O(1)$ by Assumption 2, and both $\|\mathcal{C}\|_{\max}$ and $\left\|\mathcal{C}^2\right\|_{\max}$ can be shown to be $O(1/T)$, then the desired result follows. □

Since $\partial\mathbb{P}_{NT}(\tilde{\theta})/\partial\theta = 0$, (A.64) imply that

$$\tilde{\theta} - \theta_0 = \mathcal{H}^{-1}\mathcal{S}^*(\tilde{\theta}) - \mathcal{H}^{-1}\mathcal{S}^*(\theta_0) - 0.5\mathcal{H}^{-1}\mathcal{R}(\tilde{\theta}). \tag{A.71}$$

Define

$$\mathcal{S}_{NT}^*(\theta) = \Bigg[\underbrace{\ldots, -\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\varrho^{(1)}(X_{it} - \lambda_i' f_t) f_t', \ldots,}_{1\times Nr}\underbrace{\ldots, -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\varrho^{(1)}(X_{it} - \lambda_i' f_t)\lambda_i', \ldots}_{1\times Tr}\Bigg]',$$

$\tilde{\mathcal{S}}^*(\theta) = \mathcal{S}_{NT}^*(\theta) - \mathcal{S}^*(\theta)$ and $\mathcal{D} = \mathcal{H}^{-1} - \mathcal{H}_d^{-1}$. Note that by first order conditions, $\mathcal{S}_{NT}^*(\tilde{\theta}) = 0$. As a result, we can write

$$\begin{aligned}\mathcal{H}^{-1}\mathcal{S}^*(\tilde{\theta}) =& \mathcal{H}_d^{-1}\mathcal{S}^*(\tilde{\theta}) + \mathcal{D}\mathcal{S}^*(\tilde{\theta}) = -\mathcal{H}_d^{-1}\tilde{\mathcal{S}}^*(\tilde{\theta}) + \mathcal{D}\mathcal{S}^*(\tilde{\theta}) \\ =& -\mathcal{H}_d^{-1}\tilde{\mathcal{S}}^*(\theta_0) - \mathcal{H}_d^{-1}\left(\tilde{\mathcal{S}}^*(\tilde{\theta}) - \tilde{\mathcal{S}}^*(\theta_0)\right) + \mathcal{D}\mathcal{S}^*(\tilde{\theta}) \qquad \text{(A.72)} \\ =& -\mathcal{H}_d^{-1}\tilde{\mathcal{S}}^*(\theta_0) - \mathcal{H}_d^{-1}\left(\tilde{\mathcal{S}}^*(\tilde{\theta}) - \tilde{\mathcal{S}}^*(\theta_0)\right) - \mathcal{D}\tilde{\mathcal{S}}^*(\theta_0) - \mathcal{D}\left(\tilde{\mathcal{S}}^*(\tilde{\theta}) - \tilde{\mathcal{S}}^*(\theta_0)\right). \qquad \text{(A.73)}\end{aligned}$$

Next, let $\mathcal{R}(\tilde{\theta})_j$ denote the $(j-1)r + 1$th to the jrth elements of $\mathcal{R}(\tilde{\theta})$ for $j = 1, \ldots, N+T$, and let $\bar{O}_P()$ denote a stochastic order that is uniformly in i and t.[1] Then we have

$$\begin{aligned}\mathcal{R}(\tilde{\theta})_i =& \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(3)}(*) f_t^*[f_t^{*'}(\tilde{\lambda}_i - \lambda_{0i})]^2 + \left(\frac{2}{\sqrt{NT}}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(3)}(*) f_t^* f_t^{*'}\cdot[\lambda_i^{*'}(\tilde{f}_t - f_{0t})]\right)(\tilde{\lambda}_i - \lambda_{0i}) \\ &+ \left(\frac{2}{\sqrt{NT}}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(2)}(*)\left[(\tilde{f}_t - f_{0t})f_t^{*'} + f_t^*(\tilde{f}_t - f_{0t})'\right]\right)(\tilde{\lambda}_i - \lambda_{0i}) + \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(3)}(*) f_t^*\cdot[\lambda_i^{*'}(\tilde{f}_t - f_{0t})]^2 \\ &+ \left(\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(2)}(*)(\tilde{f}_t - f_{0t})(\tilde{f}_t - f_{0t})'\right)\lambda_i^* + \bar{O}_P(1/\sqrt{T})\|\tilde{\lambda}_i - \lambda_{0i}\| + \bar{O}_P(1/T), \qquad \text{(A.74)}\end{aligned}$$

[1] For example, $Z_{it} = \bar{O}_P(1)$ means that $\max_{i\le N, t\le T}\|Z_{it}\| = O_P(1)$

$$\mathcal{R}(\tilde{\theta})_{N+t} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \bar{\varrho}_{it}^{(3)}(*)\lambda_i^* \left[\lambda_i^{*'}(\tilde{f}_t - f_{0t})\right]^2 + \left(\frac{2}{\sqrt{NT}} \sum_{i=1}^{N} \bar{\varrho}_{it}^{(3)}(*)\lambda_i^*\lambda_i^{*'}[f_t^{*'}(\tilde{\lambda}_i - \lambda_{0i})]\right)(\tilde{f}_t - f_{0t})$$
$$+ \left(\frac{2}{\sqrt{NT}} \sum_{i=1}^{N} \bar{\varrho}_{it}^{(2)}(*)\left[(\tilde{\lambda}_i - \lambda_{0i})\lambda_i^{*'} + \lambda_i^*(\tilde{\lambda}_i - \lambda_{0i})'\right]\right)(\tilde{f}_t - f_{0t}) + \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \bar{\varrho}_{it}^{(2)}(*)\lambda_i^* \left[f_t^{*'}(\tilde{\lambda}_i - \lambda_{0i})\right]^2$$
$$+ \left(\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \bar{\varrho}_{it}^{(2)}(*)(\tilde{\lambda}_i - \lambda_{0i})(\tilde{\lambda}_i - \lambda_{0i})\right) f_t^* + \bar{O}_P(1/\sqrt{T})\|\tilde{f}_t - f_{0t}\| + \bar{O}_P(1/T), \quad \text{(A.75)}$$

where $\bar{\varrho}_{it}^{(2)}(*) = \bar{\varrho}^{(2)}(\lambda_i^{*'} f_t^* - \lambda_{0i}' f_{0t})$ and $\bar{\varrho}_{it}^{(3)}(*) = \bar{\varrho}_{it}^{(3)}(\lambda_i^{*'} f_t^* - \lambda_{0i}' f_{0t})$. Write $\mathcal{D}_{j,s}$ as the $r \times r$ matrix containing the $(j-1)r+1$ to jr rows and $(s-1)r+1$ to sr columns of $\mathcal{D}$. Note that Lemma 6 and Lemma 11 imply that $\|\mathcal{H}^{-1}\mathcal{S}^*(\theta_0)\|_{\max} = \bar{O}(h^m)$. Then, from (A.71) to (A.75) we can write

$$\tilde{f}_t - f_{0t} = (\Psi_{N,t})^{-1}\frac{1}{N}\sum_{j=1}^{N} \tilde{\varrho}_{jt}^{(1)}\lambda_{0j} + \frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{s=1}^{T} \mathcal{D}_{N+t,j}\cdot \tilde{\varrho}_{js}^{(1)} \cdot f_{0s} + \frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{s=1}^{T} \mathcal{D}_{N+t,N+s}\cdot \tilde{\varrho}_{js}^{(1)}\cdot \lambda_{0j}$$
$$+ (\Psi_{N,t})^{-1}\frac{1}{N}\sum_{j=1}^{N}\left\{\tilde{\varrho}^{(1)}(X_{jt} - \tilde{\lambda}_j'\tilde{f}_t)\tilde{\lambda}_j - \tilde{\varrho}_{jt}^{(1)}\lambda_{0j}\right\} + \frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{s=1}^{T}\mathcal{D}_{N+t,j}\left\{\tilde{\varrho}^{(1)}(X_{js} - \tilde{\lambda}_j'\tilde{f}_s)\tilde{f}_s - \tilde{\varrho}_{js}^{(1)} f_{0s}\right\}$$
$$+ \frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{s=1}^{T}\mathcal{D}_{N+t,N+s}\left\{\tilde{\varrho}^{(1)}(X_{js} - \tilde{\lambda}_j'\tilde{f}_s)\tilde{\lambda}_j - \tilde{\varrho}_{js}^{(1)}\lambda_{0j}\right\}$$
$$- 0.5(\Psi_{N,t})^{-1}\mathcal{R}(\tilde{\theta})_{N+t} - 0.5\sum_{j=1}^{N}\mathcal{D}_{N+t,j}\mathcal{R}(\tilde{\theta})_j - 0.5\sum_{s=1}^{T}\mathcal{D}_{N+t,N+s}\mathcal{R}(\tilde{\theta})_{N+s} + \bar{O}(h^m). \quad \text{(A.76)}$$

Lemma 12. *Let $c_1, \ldots, c_T$ be a sequence of uniformly bounded constants. Then under Assumptions 1 and 2*

$$\frac{1}{T}\sum_{t=1}^{T} c_t(\tilde{f}_t - f_{0t}) = O_P\left(\frac{1}{Th}\right).$$

Proof. Define $d_j = \sqrt{NT}\cdot T^{-1}\sum_{t=1}^{T} c_t\mathcal{D}_{t,j}$ for $j = 1, \ldots, N+T$. Lemma 11 implies that $\max_{1\le j\le N+T}\|d_j\|$ is bounded. From (A.76), we have

$$\frac{1}{T}\sum_{t=1}^{T} c_t(\tilde{f}_t - f_{0t}) = \frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T} c_t(\Psi_{N,t})^{-1}\tilde{\varrho}_{jt}^{(1)}\lambda_{0j} + \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} d_j\tilde{\varrho}_{js}^{(1)} f_{0s} + \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} d_{N+s}\tilde{\varrho}_{js}^{(1)}\lambda_{0j}$$
$$+ \frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T} c_t(\Psi_{N,t})^{-1}\left\{\tilde{\varrho}^{(1)}(X_{jt} - \tilde{\lambda}_j'\tilde{f}_t)\tilde{\lambda}_j - \tilde{\varrho}_{jt}^{(1)}\lambda_{0j}\right\} + \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} d_j\left\{\tilde{\varrho}^{(1)}(X_{js} - \tilde{\lambda}_j'\tilde{f}_s)\tilde{f}_s - \tilde{\varrho}_{js}^{(1)} f_{0s}\right\}$$
$$+ \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} d_{N+s}\left\{\tilde{\varrho}^{(1)}(X_{js} - \tilde{\lambda}_j'\tilde{f}_s)\tilde{\lambda}_j - \tilde{\varrho}_{js}^{(1)}\lambda_{0j}\right\} - 0.5\frac{1}{T}\sum_{t=1}^{T} c_t(\Psi_{N,t})^{-1}\mathcal{R}(\tilde{\theta})_{N+t}$$
$$- 0.5\frac{1}{\sqrt{NT}}\sum_{j=1}^{N} d_j\mathcal{R}(\tilde{\theta})_j - 0.5\frac{1}{\sqrt{NT}}\sum_{s=1}^{T} d_{N+s}\mathcal{R}(\tilde{\theta})_{N+s} + \bar{O}(h^m). \quad \text{(A.77)}$$

First, by Lyapunov's CLT, it is easy to see that the first three terms on the right of (A.77) are all $O_P(1/\sqrt{NT})$.

Next, it follows from Lemma 8, (A.74), (A.75) and Assumption 2 that the last four terms on the right of (65) are all $O_P(1/L_{NT}^2)$. Finally, we will show that the remaining three terms on the right of (A.77) are all $O_P(1/(Th))$, and then the desired result follows.

Define

$$\mathbb{V}_{NT}(\theta) = \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} d_j \left\{ \tilde{\varrho}^{(1)}(X_{js} - \lambda_j' f_s) f_s - \tilde{\varrho}_{js}^{(1)} f_{0s} \right\},$$

and

$$\Delta_{NT}(\theta_a, \theta_b) = \sqrt{NT}h\left[\mathbb{V}_{NT}(\theta_a) - \mathbb{V}_{NT}(\theta_b)\right] = \underbrace{\frac{h}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{s=1}^{T} d_j \cdot \tilde{\varrho}^{(1)}(X_{js} - \lambda_{aj}' f_{as}) \cdot (f_{as} - f_{bs})}_{\Delta_{1,NT}(\theta_a,\theta_b)}$$
$$+ \underbrace{\frac{h}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{s=1}^{T} d_j \cdot \left[\tilde{\varrho}^{(1)}(X_{js} - \lambda_{aj}' f_{as}) - \tilde{\varrho}^{(1)}(X_{js} - \lambda_{bj}' f_{bs})\right] \cdot f_{bs}}_{\Delta_{2,NT}(\theta_a,\theta_b)}.$$

By Hoeffding's inequality, Lemma 2.2.1 of Van der Vaart and Wellner (1996) and the proof of Lemma 2, we can show that for $d(\theta_a, \theta_b)$ sufficiently small,

$$\|\Delta_{1,NT}(\theta_a, \theta_b)\|_{\psi_2} \lesssim T^{-1/2}\|F_a - F_b\| \lesssim d(\theta_a, \theta_b).$$

Similarly, since Lemma 6 implies that

$$h\left|\tilde{\varrho}^{(1)}(X_{js} - \lambda_{aj}' f_{as}) - \tilde{\varrho}^{(1)}(X_{js} - \lambda_{bj}' f_{bs})\right| \lesssim \left|\lambda_{aj}' f_{as} - \lambda_{bj}' f_{bs}\right|,$$

we can also show that

$$\|\Delta_{2,NT}(\theta_a, \theta_b)\|_{\psi_2} \lesssim d(\theta_a, \theta_b).$$

Thus, for $d(\theta_a, \theta_b)$ sufficiently small, we have

$$\|\Delta_{NT}(\theta_a, \theta_b)\|_{\psi_2} \le \|\Delta_{1,NT}(\theta_a, \theta_b)\|_{\psi_2} + \|\Delta_{2,NT}(\theta_a, \theta_b)\|_{\psi_2} \lesssim d(\theta_a, \theta_b).$$

Therefore, similar to the proof of Lemma 3, we can show that for sufficiently small $\delta > 0$,

$$\mathbb{E}\left[\sup_{\theta\in\Theta^M(\delta)} |\mathbb{V}_{NT}(\theta)|\right] \lesssim \frac{\delta}{L_{NT}h}. \tag{A.78}$$

It then follows from (A.78) and Lemma 8 that $\mathbb{V}_{NT}(\tilde{\theta}) = O_P(1/(L_{NT}^2 h)) = O_P(1/(Th))$, e.g., the fifth term on the right of (A.77) is $O_P(1/(Th))$. Similar results can be obtained for the fourth and sixth terms on the right of (A.77), and the thus the desired result follows. □

Lemma 13. *Under Assumptions 1 and 2, for each i we have*

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(1)}(\tilde{f}_t - f_{0t}) = O_P\left(\frac{1}{Th}\right) \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)} f_{0t}(\tilde{f}_t - f_{0t})' = O_P\left(\frac{1}{Th^2}\right).$$

Proof. To save space we only prove the second statement. Using (A.72) we can write

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)} f_{0t}(\hat{f}_t - f_{0t})' \\
&= \frac{1}{NT}\sum_{t=1}^{T}\sum_{j=1}^{N}\tilde{\varrho}_{it}^{(2)}\tilde{\varrho}_{jt}^{(1)} f_{0t}\lambda_{0j}'(\Psi_{N,t})^{-1} + \frac{1}{NT}\sum_{t=1}^{T}\sum_{j=1}^{N}\tilde{\varrho}_{it}^{(2)} f_{0t}\cdot\left\{\bar{\varrho}^{(1)}(X_{jt} - \tilde{\lambda}_j'\tilde{f}_t)\tilde{\lambda}_j' - \tilde{\varrho}_{jt}^{(1)}\lambda_{0j}'\right\}(\Psi_{N,t})^{-1} \\
&\qquad - \frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{s=1}^{T}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)} f_{0t}\tilde{f}_s'\mathcal{D}_{N+t,j}'\right)\bar{\varrho}^{(1)}(X_{js} - \tilde{\lambda}_j'\tilde{f}_s) \\
&\qquad - \frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{s=1}^{T}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)} f_{0t}\tilde{\lambda}_j'\mathcal{D}_{N+t,N+s}'\right)\bar{\varrho}^{(1)}(X_{js} - \tilde{\lambda}_j'\tilde{f}_s) \\
&\qquad + \frac{1}{2T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)} f_{0t}\mathcal{R}(\tilde{\theta})_{N+t}'(\Psi_{N,t})^{-1} + \frac{1}{2T}\sum_{j=1}^{N}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)} f_{0t}\mathcal{R}(\tilde{\theta})_j'\mathcal{D}_{N+t,j}' \\
&\qquad + \frac{1}{2T}\sum_{s=1}^{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)} f_{0t}\mathcal{R}(\tilde{\theta})_{N+s}'\mathcal{D}_{N+t,N+s}' + O(h^{m-1}). \qquad \text{(A.79)}
\end{aligned}$$

First, we can write

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{j=1}^{N}\tilde{\varrho}_{it}^{(2)}\tilde{\varrho}_{jt}^{(1)} f_{0t}\lambda_{0j}'(\Psi_{N,t})^{-1} = \frac{1}{NT}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}\tilde{\varrho}_{it}^{(1)} f_{0t}\lambda_{0j}'(\Psi_{N,t})^{-1} + \frac{1}{NT}\sum_{t=1}^{T}\sum_{j=1,j\neq i}^{N}\tilde{\varrho}_{it}^{(2)}\tilde{\varrho}_{jt}^{(1)} f_{0t}\lambda_{0j}'(\Psi_{N,t})^{-1}.$$

Since $h\varrho_{it}^{(2)}(\cdot)$ is uniformly bounded by Lemma 6, $\max_{t\leq T}\|(\Psi_{N,t})^{-1}\| = O(1)$ for large N by Assumption 2, the first term on the RHS of the above equation is $O_P((Nh)^{-1})$. Using Lyapunov's CLT and Lemma 6, the second term on the RHS of the above equation can be shown to be $O_P((NTh)^{-1/2})$. Thus, the first term on the RHS of (A.79) is $O_P((Th)^{-1})$.

Second, consider the second term on the RHS of (A.79), which can be written as

$$O_P\left(\frac{1}{Nh}\right) + \frac{1}{NT}\sum_{t=1}^{T}\sum_{j=1,j\neq i}^{N}\tilde{\varrho}_{it}^{(2)} f_{0t}\cdot\left\{\bar{\varrho}^{(1)}(X_{jt} - \tilde{\lambda}_j'\tilde{f}_t)\tilde{\lambda}_j' - \tilde{\varrho}_{jt}^{(1)}\lambda_{0j}'\right\}(\Psi_{N,t})^{-1}.$$

Similar to the proof of Lemma 12, the second term of the above expression can be shown to be $O_P(1/(Th^2))$. So the second term on the RHS of (A.79) is $O_P(1/(Th^2))$.

Next, for the third term on the RHS of (A.79), its p,qth element is given by

$$\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\chi_{i,j}\cdot\bar{\varrho}^{(1)}(X_{js} - \tilde{\lambda}_j'\tilde{f}_s)\tilde{f}_s$$

where $\chi_{i,j} = T^{-1}\sum_{t=1}^{T}\sqrt{NT}f_{0t,p}\mathcal{D}_{N+t,j,q}\tilde{\varrho}_{it}^{(2)}$, and $\mathcal{D}_{N+t,j,q}$ is the qth row of $\mathcal{D}_{N+t,j}$. Therefore,

$$\left\|\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\chi_{i,j}\,\bar{\varrho}^{(1)}(X_{js}-\tilde{\lambda}_j'\tilde{f}_s)\tilde{f}_s\right\| \leq \sqrt{\frac{1}{N}\sum_{j=1}^{N}\|\chi_{i,j}\|^2}\cdot\sqrt{\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\left[\bar{\varrho}^{(1)}(X_{js}-\tilde{\lambda}_j'\tilde{f}_s)\right]^2\|\tilde{f}_s\|^2}.$$

It is easy to show that $Th\cdot\mathbb{E}\|\chi_{i,j}\|^2 \leq \infty$ for all j, and

$$\sqrt{\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\left[\bar{\varrho}^{(1)}(X_{js}-\tilde{\lambda}_j'\tilde{f}_s)\right]^2\|\tilde{f}_s\|^2} \lesssim d(\tilde{\theta},\theta_0) = O_P(1/L_{NT})$$

by Lemma 8. So, the third term on the RHS of (A.79) is $O_P(T^{-1}h^{-1/2})$. the fourth term on the RHS of (A.79) can be shown to be $O_P(T^{-1}h^{-1/2})$ in the same way.

Finally, it follows from Lemma 8 and (A.75) that the fifth term on the RHS of (A.79) is $O_P((L_{NT})^{-2}h^{-1}) = O_P((Th)^{-1})$. The p,qth element of the sixth term on the RHS of (A.79) can be written as $(2\sqrt{NT})^{-1}\sum_{j=1}^{N}\chi_{i,j}\mathcal{R}(\tilde{\theta})_j$, which is bounded by

$$\frac{\sqrt{N}}{2\sqrt{T}}\sqrt{\frac{1}{N}\sum_{j=1}^{N}\|\chi_{i,j}\|^2}\sqrt{\frac{1}{N}\sum_{j=1}^{N}\|\mathcal{R}(\tilde{\theta})_j\|^2} = O_P((Th)^{-1/2})O_P(L_{NT}^{-2}) = O_P(T^{-3/2}h^{-1/2}).$$

The same bound for the seventh term on the RHS of (A.79) can be obtained using the same argument. Thus, combining the above results, we get

$$\frac{1}{T}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(2)}f_{0t}(\tilde{f}_t-f_{0t})' = O_P\left(\frac{1}{Th^2}\right).$$

□

Proof of Theorem 4:

Proof. From the expansion in the proof of Lemma 10,

$$\Phi_{T,i}(\tilde{\lambda}_i-\lambda_{0i}) = -\frac{1}{T}\sum_{t=1}^{T}\bar{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'\tilde{f}_t)\tilde{f}_t + \frac{1}{T}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(1)}f_{0t} + \frac{1}{T}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(1)}(\tilde{f}_t-f_{0t})$$
$$-\frac{1}{T}\sum_{t=1}^{T}\bar{\varrho}_{it}^{(2)}f_{0t}(\tilde{f}_t-f_{0t})'\lambda_{0i} + O_P\left(T^{-1}\|\tilde{F}-F_0\|^2\right) + o_P(\|\tilde{\lambda}_i-\lambda_{0i}\|).$$

It then follows from Lemma 6, Lemma 8 and Lemma 12 that

$$\Phi_{T,i}(\tilde{\lambda}_i-\lambda_{0i}) = -\frac{1}{T}\sum_{t=1}^{T}\bar{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'\tilde{f}_t)\tilde{f}_t + O(h^m) + O_P\left(\frac{1}{Th}\right) + o_P(\|\tilde{\lambda}_i-\lambda_{0i}\|).$$

Note that

$$
\begin{aligned}
-\frac{1}{T}\sum_{t=1}^{T}\bar{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'\tilde{f}_t)\tilde{f}_t &= \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}^{(1)}(X_{it}-\tilde{\lambda}_i'\tilde{f}_t)\tilde{f}_t \\
&= \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(1)}\tilde{f}_t - \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}\cdot(\tilde{\lambda}_i'\tilde{f}_t-\lambda_{0i}'f_{0t})\tilde{f}_t + 0.5\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(3)}(*)(\tilde{\lambda}_i'\tilde{f}_t-\lambda_{0i}'f_{0t})^2\tilde{f}_t,
\end{aligned}
$$

where $\tilde{\varrho}_{it}^{(3)}(*)=\tilde{\varrho}_{it}^{(3)}(c_{it}^*)$ and c_{it}^* is between $\lambda_{0i}'f_{0t}$ and $\tilde{\lambda}_i'\tilde{f}_t$.

First, by Lemma 13 we have

$$
\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(1)}\tilde{f}_t = \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(1)}f_{0t} + \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(1)}(\tilde{f}_t-f_{0t}) = \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(1)}f_{0t} + O_P\left(\frac{1}{Th}\right).
$$

Second,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}\cdot(\tilde{\lambda}_i'\tilde{f}_t-\lambda_{0i}'f_{0t})\tilde{f}_t &= \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}\tilde{f}_t\cdot(\tilde{f}_t-f_{0t})'\tilde{\lambda}_i + \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}\tilde{f}_tf_{0t}'\cdot(\tilde{\lambda}_i-\lambda_{0i}) \\
= \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}f_{0t}\cdot(\tilde{f}_t-f_{0t})'\tilde{\lambda}_i &+ \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}(\tilde{f}_t-f_{0t})\cdot(\tilde{f}_t-f_{0t})'\tilde{\lambda}_i + \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}f_{0t}f_{0t}'\cdot(\tilde{\lambda}_i-\lambda_{0i}) \\
&+ \frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}(\tilde{f}_t-f_{0t})f_{0t}'\cdot(\tilde{\lambda}_i-\lambda_{0i}). \quad \text{(A.80)}
\end{aligned}
$$

It then follows from Lemma 8 that the second term on the RHS of (A.80) is $O_P((Th)^{-1})$, and Lemma 13 implies that the first term is $O_P(T^{-1}h^{-2})$. It is easy to show that the last two terms on the right of (A.80) are both $o_P(\|\tilde{\lambda}_i-\lambda_{0i}\|)$. Thus,

$$
\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(2)}\cdot(\tilde{\lambda}_i'\tilde{f}_t-\lambda_{0i}'f_{0t})\tilde{f}_t = O_P(T^{-1}h^{-2}) + o_P(\|\tilde{\lambda}_i-\lambda_{0i}\|). \quad \text{(A.81)}
$$

Next, it is also easy to show that

$$
\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(3)}(*)(\tilde{\lambda}_i'\tilde{f}_t-\lambda_{0i}'f_{0t})^2\tilde{f}_t\right\| \leq K_1\|\tilde{\lambda}_i-\lambda_{0i}\|^2\frac{1}{T}\sum_{t=1}^{T}|\tilde{\varrho}_{it}^{(3)}(*)| + \frac{K_2}{T}\sum_{t=1}^{T}|\tilde{\varrho}_{it}^{(3)}(*)|\cdot\|\tilde{f}_t-f_{0t}\|^2,
$$

Therefore, from Lemma 6, Lemma 8 and Lemma 10 we have

$$
\frac{1}{T}\sum_{t=1}^{T}\tilde{\varrho}_{it}^{(3)}(*)(\tilde{\lambda}_i'\tilde{f}_t-\lambda_{0i}'f_{0t})^2\tilde{f}_t = O_P(\|\tilde{\lambda}_i-\lambda_{0i}\|)\cdot O_P(T^{-1/2}h^{-3}) + O_P(T^{-1}h^{-2}),
$$

which is $o_P(\|\tilde{\lambda}_i-\lambda_{0i}\|)+O_P(T^{-1}h^{-2})$ under the assumption that $\sqrt{T}h^3\to\infty$.

Finally, combining all the results above we get

$$\Phi_{T,i}(\tilde{\lambda}_i - \lambda_{0i}) = \frac{1}{T}\sum_{t=1}^{T} \tilde{\varrho}_{it}^{(1)} f_{0t} + o_P(\|\tilde{\lambda}_i - \lambda_{0i}\|) + O_P\left(\frac{1}{Th^2}\right) + O(m), \tag{A.82}$$

and from Lemma 6 it is easy to show that:

$$\frac{1}{T}\sum_{t=1}^{T} \bar{\varrho}_{it}^{(2)} f_{0t} f_{0t}' \to \Phi_i > 0 \quad \text{and} \quad \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \tilde{\varrho}_{it}^{(1)} f_{0t} \xrightarrow{d} \mathcal{N}(0, \tau(1-\tau)\mathbb{I}_r). \tag{A.83}$$

Since our assumption implies that $\sqrt{T}h^2 \to \infty$ and $\sqrt{T}h^m \to 0$, the desired results follow from (A.82) and (A.83). $\square$

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Chen, M., I. Fernández-Val, and M. Weidner (2018). Nonlinear factor models for network and panel data. *Working paper, UCL.*

Galvao, A. F. and K. Kato (2016). Smoothed quantile regression for panel data. *Journal of Econometrics 193*(1), 92–112.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica*, 1327–1351.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Van der Vaart, A. and J. Wellner (1996). *Weak convergence and empirical processes.* Springer, New York.